

**Cross-Modality  
Semantic Integration and  
Robust Interpretation of  
Multimodal User Interactions**

**HUI, Pui Yu**

A Thesis Submitted in Partial Fulfilment  
of the Requirements for the Degree of  
Doctor of Philosophy  
in  
Systems Engineering and Engineering Management

The Chinese University of Hong Kong  
September 2010

UMI Number: 3483863

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3483863

Copyright 2011 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346



Thesis/Assessment Committee

Professor Wai Lam (Chair)

Professor Helen M. Meng (Thesis Supervisor)

Professor Hong Cheng (Committee Member)

Professor James Landay (External Examiner)

Abstract of thesis entitled:

Cross-Modality Semantic Integration and Robust Interpretation of Multimodal User Interactions

Submitted by HUI, Pui Yu

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in September 2010

Multimodal systems can represent and manipulate semantics from different human communication modalities at different levels of abstraction, in which multimodal integration is required to integrate the semantics from two or more modalities and generate an interpretable output for further processing. In this work, we develop a framework pertaining to automatic cross-modality semantic integration of multimodal user interactions using speech and pen gestures. It begins by generating partial interpretations for each input event as a ranked list of hypothesized semantics. We devise a *cross-modality semantic integration procedure* to align the pair of hypothesis lists between every speech input event and every pen input event in a multimodal expression. This is achieved by the Viterbi alignment that enforces the temporal ordering and semantic compatibility constraints of aligned events. The alignment enables generation of a *unimodal paraphrase* that is semantically equivalent to the original multimodal expression. Our experiments are based on a multimodal corpus in the navigation domain. Application of the integration procedure to manual transcripts shows that correct unimodal paraphrases are generated for around 96% of the multimodal inquiries in the test set. However, if we replace this

with automatic speech and pen recognition transcripts, the performance drops to around 53% of the test set. In order to address this issue, we devised the *hypothesis rescoring procedure* that evaluates all candidates of cross-modality integration derived from multiple recognition hypotheses from each modality. The rescoring function incorporates the integration score,  $N$ -best purity of recognized spoken locative references (SLRs), as well as distances between coordinates of recognized pen gestures and their interpreted icons on the map. Application of *cross-modality hypothesis rescoring* improved the performance to generate correct unimodal paraphrases for over 72% of the multimodal inquiries of the test set.

We have also performed a latent semantic modeling (LSM) for interpreting multimodal user input consisting of speech and pen gestures. Each modality of a multimodal input carries semantics related to a domain-specific task goal (TG). Each input is annotated manually with a TG based on the semantics. Multimodal input usually has a simpler syntactic structure and different order of semantic constituents from unimodal input. Therefore, we proposed to use LSM to derive the latent semantics from the multimodal inputs. In order to achieve this, we characterized the cross-modal integration pattern as 3-tuple multimodal terms taking into account SLR, pen gesture type and their temporal relation. The correlation term matrix is then decomposed using singular value decomposition (SVD) to derive the latent semantics automatically. TG inference on disjoint test set based on the latent semantics achieves accurate performance for 99% of the multimodal inquiries.

多模態系統 (multimodal system) 能夠表達和操作不同層次的抽象概念的各種人類溝通方式中的語意，當中，需要多模態語意集成 (multimodal integration) 去整合兩個或以上的模態的語意，並輸出可解釋的結果作進一步的處理。在這項工作中，我們制定一個能應用於語音及筆觸輸入平台的跨模態自動語意整合 (automatic cross-modality semantic integration) 框架。擬議的框架首先為每個輸入事件產生順序的部分解釋，再利用我們所制定的一個跨模式的語意集成過程，以維特比對齊算法 (Viterbi alignment) 根據時間順序 (temporal ordering) 及語意兼容性 (semantic compatibility) 來把每個多模態輸入中的語音及筆觸事件作對應，再生成一個相當於原有的多模態輸入的單模態釋義 (unimodal paraphrase)。我們的實驗是基於一個多模態城市導航語料庫。於手工謄本 (manual transcripts) 上應用跨模態自動語意整合 (cross-modality semantic integration) 顯示能正確為大約96%的測試集中的多模態句子生成單模態釋義 (unimodal paraphrases)。然而，當我們把框架套用到自動語音及筆觸識別的結果上，便發現測試集的正確率下降至52.5%。為了解決這個問題，我們設計了假設重新記分程序 (hypothesis rescoring procedure)，用於重新評估所有跨模態自動語意整合中的候選人的評分。該重新評分程序利用整合評分 (integration score)、語音地點指示詞 (spoken locative reference) 的評分及筆觸輸入與地圖上地點坐標之間的距離評分作重新評分。假設重新記分程序 (hypothesis rescoring procedure) 改善了實驗結果，能正確地為72.7%的測試集中的多模態句子生成單模態釋義。

我們也制定了應用在語音及筆觸輸入上的潛在語意模型 (latent semantic modeling) 框架。每項模態也有跟領域目標 (task goal) 有關的語意，而每個輸入也手工註釋了對應的領域目標 (task goal)。多模態輸入通常較單式輸入有比較簡單的語法結構，而兩者語意成分的順序也有分別。因此，我們建議採用潛在語意模型以推導出多模態輸入中的潛在語意。為了達到這個目標，我們把跨模態輸入格式 (cross-modal integration pattern) 以三元組多模態詞彙 (3-tuple multimodal terms) 的方式表達，以顯示出語音地點指示詞、筆觸輸入類型及他們之間的時間關係。我們再利用奇異值分解

(Singular Value Decomposition)分解查詢－多模態詞彙矩陣，然後自動推導出潛在語意。最後，發現以潛在語意能成功為99%測試集中的多模態句子推導其領域目標(task goal)。

# Acknowledgements

I would like to thank my supervisor, Professor Helen Meng, for her guidance throughout this research project. Her comments and feedback are invaluable to my research and thesis writing. I also thank Helen for her experience sharing, all the training and opportunities she gave me. She talked with me when I had emotional upset, discussed with me when there was a puzzle to me and gave me chances to learn from the experts in this research area. She puts trust in us, gives us freedom to choose our research topics and ways to achieve them. She also provides us with excellent computing resources and facilities so that we can focus on our research and schoolwork.

I also want to thank my entire thesis committee, Professor Helen Meng, Professor Wai Lam, Professor Hong Cheng and Professor James Landay, for their time, effort and valuable advice. I would like to thank Professor Irwin King for his encouragement and advice in event coordination over the years. Besides, I had the good fortune to gain knowledge from Professor Frank Soong, the late Dr. Jian-Lai Zhou, Dr. Ye Tian and Dr. Chao Huang from Microsoft Research Asia (MSRA), they spent much time and patience to teach me and provided me a lot of helpful suggestions in Mandarin speech recognition.

Members in MSRA and Human-Computer Communications Laboratory (HCCL) helped me a lot too. I want to thank Professor Frank Soong, Chao Huang, Jianlai Zhou, Ye Tian, Honghui Dong, Fazhou Wu, Qiang Fu, Jiali You, Heng Kang, Min Lai, Lijuan Yang, Peng Liu, Minho Jin, Yu Shi, Yining

Chen, Yvonne Lee, Yong Zhao, Jessica Zhou, Zhijie Yan and Xi Zhou from the Speech Group of MSRA for their participation in the multimodal data collection; Qiben Wang from the Incubation Group of MSRA for being my lovely room-mate during the internship. I also want to thank Gangi Reddy from IIT, Jay Young and Sarah Kim from Yonsei University for their work and contribution in pen gesture recognition during their internship in HCCL; Jessica Zhou for her help in perplexity calculation; Simon Wong for the development of multimodal dialog system on CUHK navigation; Jackie Cao for the development of a visualizer for the rendition of collected multimodal input; Alissa Harrison for providing me some useful programs and her help in laboratory resources management; Dr. Wai-kit Lo for his advices on research issues; and all the team members for their time in chitchat.

Friends, including Agatha, Amelia, Carol, NGok, Eliza, Steven, Erica, Hoi-ka, Leo, Rita, the late Dr. Anthony Yeung and Professor Carole Hoyan Yeung, support and bring happiness to me all the time. Thanks for all the joyful time they gave me.

Staffs from the General Office and Technical Team of our Department are very nice and thank you very much for their support in these years.

Finally, I would like to thank KT for his advices, encouragements and reminders throughout these years; my brother Ray and his best friend Loret for their wordless support to all the decisions I have made; and most importantly, my Mum Anne and my husband Simon for their endless love and the confidence they continuously give me.

This work is partially supported by a grant from the HKSAR Government Research Grants Council (Project No. 415609). This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.



# Contents

<b>Abstract</b>	<b>3</b>
<b>Acknowledgements</b>	<b>7</b>
<b>1 Introduction</b>	<b>30</b>
1.1 Challenges . . . . .	34
1.1.1 Recognition of User Input in Different Modalities . . .	34
1.1.2 Message Fission . . . . .	35
1.1.3 Cross-modality Integration . . . . .	36
1.1.4 High-dimensionality . . . . .	36
1.2 Thesis Goals . . . . .	37
1.3 Thesis Organization . . . . .	38
<b>2 Related Work</b>	<b>40</b>
2.1 Variety of Modalities . . . . .	40
2.2 Relationships across Modalities . . . . .	43
2.2.1 Speech-Gesture Correlation . . . . .	43
2.2.2 Cognitive Status and Form of References . . . . .	44
2.2.3 Integration and Synchronization of Input Modalities . .	46
2.2.4 CARE Properties . . . . .	47
2.3 Semantic Interpretation/Integration Methods . . . . .	49
2.4 Evaluation Methods . . . . .	53

2.4.1	Component-level Evaluation . . . . .	53
2.4.2	System-level Evaluation . . . . .	53
2.5	Chapter Summary . . . . .	55
<b>3</b>	<b>Multimodal Corpus</b>	<b>56</b>
3.1	Information Domain . . . . .	56
3.2	Data Collection Procedures . . . . .	60
3.3	Data Collection Setup . . . . .	63
3.4	Corpus Statistics . . . . .	65
3.5	Manual Annotation of Cross-Modality Pairings . . . . .	67
3.6	Chapter Summary . . . . .	72
<b>4</b>	<b>Unimodal and Cross-modal Characterizations</b>	<b>74</b>
4.1	Characterization of Spoken Inputs . . . . .	76
4.2	Procedure for Interpreting Spoken Locative References . . . . .	78
4.3	Characterization of Pen Inputs . . . . .	79
4.4	Interpreting Pen Inputs . . . . .	80
4.5	Temporal Relationships . . . . .	83
4.6	Cross-Modal Integration Patterns . . . . .	89
4.7	Chapter Summary . . . . .	94
<b>5</b>	<b>Cross-Modality Semantic Integration</b>	<b>95</b>
5.1	Cross-Modality Integration on Perfect Transcriptions . . . . .	96
5.1.1	Enforcing Temporal Order . . . . .	97
5.1.2	Enforcing Semantic Compatibility . . . . .	99
5.1.3	Identifying Intended Locations . . . . .	100
5.1.4	Evaluating the Cross-Modality Integration Procedure . . . . .	104
5.2	Analytical Comparison between Parallel Multimodal and Unimodal Expressions . . . . .	107
5.2.1	Language Modeling . . . . .	109

5.2.2	Data Analysis . . . . .	110
5.3	Cross-Modality Integration on Imperfect Transcriptions . . . .	117
5.3.1	Transcribing the Spoken Inputs . . . . .	118
5.3.2	Recognizing the Pen Inputs . . . . .	119
5.3.3	Evaluating the Cross-Modality Integration Procedure .	122
5.4	Chapter Summary . . . . .	124
<b>6</b>	<b>Hypothesis Rescoring for Robustness towards Imperfect Tran- scriptions</b>	<b>132</b>
6.1	Pruning and Scoring the Recognized Spoken Inputs . . . . .	133
6.2	Filtering and Scoring the Recognized Pen Inputs . . . . .	136
6.3	Pruning and Scoring Cross-Modality Integrations . . . . .	140
6.4	Rescoring Cross-Modality Integrations . . . . .	142
6.5	Evaluating the Rescoring Procedure . . . . .	147
6.6	Chapter Summary . . . . .	149
<b>7</b>	<b>Latent Semantic Analysis for Multimodal User Input with Speech and Pen Gestures</b>	<b>155</b>
7.1	Latent Semantic Modeling of Cross-modal Integration Patterns	156
7.1.1	Association Matrices . . . . .	157
7.2	Task Goal Inference . . . . .	160
7.2.1	Performance Baseline using Vector-Space Model . . . .	160
7.2.2	Performance Evaluation . . . . .	162
7.2.3	Analysis of the Re-constructed Space for Identification of Key Terms . . . . .	163
7.3	Chapter Summary . . . . .	165
<b>8</b>	<b>Latent Semantic Analysis for Task Goal Inference</b>	<b>174</b>
8.1	Latent Semantic Modeling for Task Goal Inference . . . . .	175
8.1.1	Association Matrices . . . . .	175

8.1.2	Relating Task Goals with Latent Semantics . . . . .	177
8.2	Task Goal Inference . . . . .	181
8.2.1	Performance Baseline using Vector-Space Model . . . . .	181
8.2.2	Optimization of $R'$ . . . . .	182
8.2.3	Performance Evaluation . . . . .	186
8.3	Task Goal Inference with Spoken Terms Regularization . . . . .	187
8.3.1	Spoken Terms Regularization . . . . .	187
8.3.2	Performance Baseline using Vector-Space Model . . . . .	188
8.3.3	Optimization of $R'$ . . . . .	189
8.3.4	Performance Evaluation . . . . .	191
8.4	Analysis of the Latent Semantic Space for Task Goal Inference	194
8.4.1	Sub-categorization of task goals . . . . .	194
8.4.2	Capturing key terms for task goals . . . . .	201
8.4.3	Generalizing across related multimodal terms . . . . .	201
8.5	Error Analysis of the Latent Semantic Space for Task Goal Inference . . . . .	207
8.6	Chapter Summary . . . . .	208
<b>9</b>	<b>Conclusions and Future Work</b>	<b>211</b>
9.1	Thesis Summary . . . . .	211
9.2	Contributions . . . . .	215
9.3	Future Work . . . . .	217
<b>A</b>	<b>A Survey on Information Categories</b>	<b>220</b>
<b>B</b>	<b>User Tasks for Data Collection</b>	<b>223</b>
<b>C</b>	<b>An Instruction Provided by a Subject</b>	<b>237</b>
<b>D</b>	<b>An Illustrative Example on the Normalized Cost <math>C_S(S_r, N)</math> for the Recognized SLR <math>S_r</math></b>	<b>248</b>

<b>E</b>	<b>An Illustrative Example on the Hypothesis Rescoring Procedure</b>	<b>254</b>
<b>F</b>	<b>Significance Tests</b>	<b>272</b>
F.1	Cross-modality hypotheses rescoring of the First Best Recognized Speech Hypotheses and $M$ -Best Pen Recognition Outputs ( $M = 4$ ) . . . . .	273
F.2	Cross-modality hypotheses rescoring of the $N$ -Best Speech Recognition Hypotheses ( $N = 100$ ) and First Best Pen Recognition Outputs . . . . .	277
F.3	Cross-modality hypotheses rescoring of the $N$ -Best Speech Recognition Hypotheses ( $N = 100$ ) and $M$ -Best Pen Recognition Outputs ( $M = 4$ ) . . . . .	281
F.4	Improvements in the Integration Accuracy brought about by Cross-Modality Hypotheses Rescoring in the Presence of Speech Recognition Errors . . . . .	283
F.5	Improvements in the Integration Accuracy brought about by Cross-Modality Hypotheses Rescoring in the Presence of Pen Recognition Errors . . . . .	286
F.6	Improvements in the Integration Accuracy brought about by Cross-Modality Hypotheses Rescoring in the Presence of both Speech and Pen Recognition Errors . . . . .	288
<b>G</b>	<b>Abbreviations</b>	<b>292</b>
	<b>Bibliography</b>	<b>294</b>

# List of Figures

1.1	An illustration of human-human communication with multiple modalities. . . . .	30
1.2	An illustration of complementarity between speech and hand movements. In this example, the person is narrating a story. He says, “ <i>she chases him out</i> ” with hand movements that appear to swing an object through the air. The hand movements show the swinging of a weapon while speech conveys the action of chasing. This example is borrowed from [1]. . . . .	32
1.3	An illustration of redundancy between speech and arm movements. In this example, the person says, “ <i>he bends it way back</i> ” with arm movements that appear to grip something and pull it back. The arm movements exhibit the same semantic content as presented in speech. This example is borrowed from [1]. . .	32
2.1	Semantic-level integration (top) vs. feature-level integration (bottom). . . . .	50
3.1	The map of the Beijing City. The coverage of the maps we downloaded from the Internet is highlighted in blue. They are Haidian District, Xicheng District, Dongcheng District, Chaoyang District, Fentai District, Xuanwu District and Chongwen District.	58

3.2	A map downloaded from the Internet. The numbers highlight some examples of location icons for the LOC_TYPE of TRANSPORTATION, including subtypes of (1) <i>railroad</i> , (2) <i>train station</i> , (3) <i>elevated highway</i> , (4) <i>railway station</i> , (5) <i>intersection</i> , (6) <i>bridge</i> and (7) <i>highway</i> ; location icons for LOC_TYPE of LEISURE FACILITIES, including subtypes of (8) <i>exhibition center</i> , (9) <i>green area</i> and (10) <i>museum</i> ; and location icon of PUBLIC FACILITIES AND SERVICES, including (11) <i>temple</i> . . . . .	59
3.3	Data collection interface of the Pocket PC, augmented with soft buttons for logging functions (START/STOP) and loading the NEXT map. The numbers highlight some examples of location icons: (1) subject's current location (i.e. the red cross); (2) a university; (3) a road and (4) a hospital. . . . .	65
3.4	The distribution of multimodal inquiries with different numbers of non-spurious SLRs and pen gestures (i.e. manual transcriptions) in the training set. The number inside a bubble is the number of multimodal inquiries with that particular number of SLRs and pen gestures. . . . .	69
3.5	An illustration on the nature of the annotator. . . . .	70

3.6	An illustration of the process of manual annotation of cross-modality pairings. The input contains three SLRs which correspond to two locations on the map, together with two circling pen gestures. The locations in the boxes are the possible interpretations of an input event. The boxes with the same border are paired up and the underlined locations are the matched locations between the paired, cross-modal events. The indices of SLRs and pen gestures begin at zero. (1,1) means that the first SLR is aligned with the first pen gestures. In this example, the second SLR “here” does not align with any pen gesture so its pairing is labeled as (2,).	73
4.1	Distribution of SLRs according to the types of referent in the training set.	77
4.2	Distribution of the types of SLRs according to their numeric features in the training set.	78
4.3	Distribution of lag times between end of speech and onset of pen (i.e. speech precedes) in sequential inputs.	88
4.4	Distribution of lag times between end of pen and onset of speech (i.e. pen precedes) in sequential inputs.	88
5.1	The cross-modality integration procedure. Each input event (a spoken locative reference or a pen gesture such as POINT/CIRCLE/STROKE) in each modality produces a list of hypothesized locations. There are aligned across modalities by the Viterbi algorithm while incorporating semantic compatibility and temporal order. $S_r[N]$ is the $N$ -best recognition hypothesis of the $r^{th}$ SLR and $P_q[M, K_{q,M}]$ is the $M$ -best recognition hypothesis of the $q^{th}$ pen gesture instance with $K_{q,M}$ hypothesized locations.	98



5.2	An Illustrative Example on the Viterbi Alignment Algorithm. The arrows are the back pointer of $(S_r, P_q)$ , which has minimum cumulative cost $C_A(S_r, P_q)$ . . . . .	103
5.3	Speech recognition performance (character accuracies) across subjects in the training set of the multimodal corpus. . . . .	120
5.4	An illustration of convex hull in a circle. . . . .	121
5.5	A circle formed by three points, $x$ , $y$ and $z$ and the radius of curvature. . . . .	121
5.6	Performance of the cross-modality integration (CMI) in the training and test sets. . . . .	123
5.7	Performance of the cross-modality integration (CMI) in the training and test sets. . . . .	125
5.8	Plot of the relation between the performance of cross-modality integration procedure and overall speech recognition accuracy. . . . .	126
6.1	The system architecture of cross-modality integration with hy- pothesis rescoring, which can be the front-end multimodal pro- cessing framework for an existing unimodal dialog system. . . . .	133
6.2	An illustration on the comparison between the four corners (i.e. maximum and minimum values of $x$ and maximum and mini- mum values of $y$ ) of two circles. . . . .	137
6.3	Performance of cross-modality integration (CMI) in the training and test sets. . . . .	148
7.1	Similarity between vector $j_a$ and $b_n$ captured by the cosine of the angle $\theta$ between them. The angle $\theta$ bewteen the two vectors is 0, corresponding to maximal similarity. . . . .	161
7.2	A plot of term weight from matrix $\hat{B}$ against lexical and multi- modal terms ( $M = 881$ ) for the task goal BUS INFORMATION. . . . .	163

7.3	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal CHOICE OF VEHICLE.	166
7.4	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal MAP COMMANDS.	167
7.5	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal OPENING HOURS.	168
7.6	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal RAILWAY INFORMATION.	169
7.7	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal ROUTE FINDING.	170
7.8	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal TIME CONSTRAINT.	171
7.9	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal TRANSPORTATION COSTS.	172
7.10	A plot of term weight from matrix $\hat{B}$ against lexical and multimodal terms ( $M = 881$ ) for the task goal TRAVEL TIME.	173
8.1	A plot of the cumulative percentage of the singular values against the order of SVD approximation.	183
8.2	A plot of task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation.	184
8.3	A plot of task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation for the optimization of $R'$ .	185
8.4	Performance of task goal inference for each of the nine task goals in the application domain. Results are based on the latent space with 406 dimensions.	186

8.5	A plot of the cumulative percentage of the singular values against the order of SVD approximation. . . . .	190
8.6	A plot of the task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation. . . . .	191
8.7	A plot of the task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation for the optimization of $R'$ . . . . .	192
8.8	Performance of task goal inference for each of the nine task goals in the application domain. Results are based on the latent space with 263 dimensions. . . . .	193
8.9	Percentage of multimodal inputs that belong to different latent semantic categories, within the task goal BUS INFORMATION. The numbers inside the bars are the labels (indexed by $r$ ) of the latent semantic categories. . . . .	194
8.10	Percentage of multimodal inputs that belong to different latent semantic categories, within the task goal OPENING HOURS. The numbers inside the bars are the labels (indexed by $r$ ) of the latent semantic categories. . . . .	197
8.11	A plot of term weight from matrix $\hat{G}$ against lexical and multimodal terms ( $M = 477$ ) for the task goal MAP COMMANDS. . . . .	202
8.12	A plot of term weight from matrix $\hat{G}$ against lexical and multimodal terms ( $M = 477$ ) for the task goal ROUTE FINDING. . . . .	203

# List of Tables

2.1	Correlation between linguistic form and status in Givenness Hierarchy where N stands for noun and $\emptyset$ means zero. $\emptyset$ N (in the last column) indicates the use of a noun only and no article is required in Chinese (e.g. <i>dog</i> ). The arrows indicate that the statuses are ordered from most restrictive to least restrictive with respect to the possible referents they include. . . . .	44
3.1	A complete listing of location types and their corresponding subtypes. . . . .	57
3.2	An illustrative example for multimodal data collection with speech ( <i>S</i> ) and pen gestures ( <i>P</i> ). Translations are italicized. . . . .	61
3.3	Examples given to subjects as illustrations of the use of different pen gestures. . . . .	64
3.4	Example of logged data for multimodal input based on Table 3.2. Explanations are in italics. . . . .	66
3.5	An example of manual transcription output. All the SLRs and pen gestures are indicated with $\langle \dots \rangle$ . . . . .	67
3.6	Details of the multimodal corpus collected. . . . .	68
4.1	Overview statistics of the Multimodal Corpus. Translations are italicized. . . . .	75
4.2	Illustrations of the usages of different pen gesture types. . . . .	81

4.3	Illustrations of the usages of multiple pen gestures. . . . .	82
4.4	An illustration of the procedure for hypothesis lists generation in the speech and pen modalities respectively. Translations are italicized. Distance labeled with “-1” means the pen gesture is triggered inside the area of icon/label of that location. . . . .	85
4.5	Nine logically possible temporal overlap patterns between speech and pen gestures for simultaneous inputs. . . . .	87
4.6	Two temporal patterns between speech and pen gestures for sequential inputs. . . . .	87
4.7	Percentage of simultaneous and sequential temporal patterns for all 23 subjects. Average consistency of user’s dominant integration pattern is 82.6%. . . . .	90
4.8	Statistics of cross-modal integration patterns in the training set. There are altogether 2480 multimodal terms (count by token) in total. Among them, 2261 contain both SLR and pen gesture, 181 contain only SLRs and 38 of them contain only pen gestures.	92
4.9	Examples on 3-tuple multimodal term annotation with speech ( <i>S</i> ) and pen gesture ( <i>P</i> ). Translations are italicized. . . . .	93
5.1	Details of the Viterbi Alignment Algorithm. . . . .	102
5.2	An example illustrating the unimodal paraphrases generated from the multimodal expressions from two dialog turns. MM1 and MM2 are the multimodal expressions from dialog turns 1 and 2. UM1 and UM2 are the unimodal paraphrases generated from MM1 and MM2 respectively. Translations are italicized. . . . .	105
5.3	An example on the incorrect alignment due to the presence of redundant SLRs (i.e. four “ <i>this</i> ” and one “ <i>these four places</i> ”) in the speech input. . . . .	107

5.4	Parallel multimodal and unimodal corpora statistics. The difference in the total number of words in multimodal input and unimodal paraphrase may due to the use of plural and aggregated references in the multimodal input. . . . .	108
5.5	Comparisons in perplexities between the parallel multimodal (MM) and unimodal (UM) inputs. The difference in the number of words is less than expected due to the diversity of Chinese measure words and contextual phrases mentioned. . . . .	109
5.6	Comparison of per-utterance perplexities between the multimodal inputs and their unimodal paraphrases. . . . .	110
5.7	Illustrative examples from the testing data subset with ( $PP_{MM} = PP_{UM}$ ). . . . .	112
5.8	Illustrative examples from the testing data subset with ( $PP_{MM} = PP_{UM}$ ). . . . .	114
5.9	Overall statistics of different categories found by comparison of per-utterance perplexities between the multimodal inputs and their unimodal paraphrases. . . . .	115
5.10	Examples illustrating perplexity reduction in different cases. Translations are italicized. . . . .	116
5.11	Illustrative examples on the recognition errors of circle and stroke.	122
5.12	Performance of the cross-modality integration, measured in terms of percentage of correctly aligned expressions in the training and test sets. . . . .	128
5.13	Examples on the correct integration with the present of SLR recognition error. . . . .	130

5.14	Performance of cross-modality integration, measured in terms of the percentage of correctly aligned expressions in the training and test sets based on the Viterbi Alignment in Chapter 5 and Table 5.1 in Section 5.1. . . . .	131
6.1	An example showing the normalized cost of each recognized SLR based on Equation 6.1 for the $N$ -best ( $N = 100$ ) recognition hypotheses. . . . .	136
6.2	An illustrative example for the calculation of normalized cost for the top-scoring (i.e. $m = 1$ ) recognized pen gesture. In this example, the multimodal is transcribed as a sequence of one pen gesture (i.e. $Q = 1$ ) and there are six hypothesized locations in total (i.e. $K_{q,1} = 6$ ) . . . . .	139
6.3	An illustrative example of the pruning mechanism for candidates for cross-modality integrations. The first SLR of the top-scoring speech recognition hypothesis is the abbreviated name of “ <i>Beijing Medical University</i> ” while the first SLR of the second-best speech recognition hypothesis is the abbreviated name of “ <i>Beijing University of Post and Telecommunications.</i> ” . . . . .	141

6.4	An example illustrating the hypothesis rescoring process of based on the $N$ -best speech recognition hypotheses ( $N = 100$ ) listed in Table 6.1. The second SLR, <i>these places</i> , should have NUM=plural, which can be aligned with more than one pen gestures. Another possibility of the second SLR is <i>here</i> , which should have NUM=nil and can be aligned with any number of pen gestures. All the five pen gestures incur no cost because their coordinates coincide with the respect icons/labels. Each candidate for cross-modality integration is rescored and then the updated rank is shown for each candidate. The 98 <sup>th</sup> hypothesis pair ranked top after rescoring. . . . .	147
6.5	Performance of cross-modality integration, measured in terms of the percentage of correctly aligned expressions in the training and test sets. Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from (1) top-scoring hypotheses from speech and pen to top-scoring hypotheses from speech and $M$ -best hypotheses from pen; (2) top-scoring hypotheses from speech and pen to $N$ -best hypotheses from speech and top-scoring hypotheses from pen; (3) top-scoring hypotheses from speech and pen to $N$ -best hypotheses from speech and $M$ -best hypotheses from pen ( $\alpha = 0.01$ , two-tailed $z$ -test) as shown in Appendix F. . . .	150
6.6	Detailed performance statistics of the test set. Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant in the presence of speech and/or pen recognition errors ( $\alpha = 0.01$ , two-tailed $z$ -test) as shown in Appendices F.4, F.5 and F.6. . . . .	151



6.7	Examples on the correct integration with the present of SLR and/or pen recognition error. . . . .	153
6.8	An example on the incorrect alignment due to the presence of NUM=plural (from “ <i>these places</i> ”) and missing of timing information during integration. Since NUM feature of <i>these places</i> is plural, which can align with <i>more than one pen gestures</i> without a specific number, our framework align one of the pen gesture to the anaphora (i.e. the second “ <i>here</i> ”) in the spoken input. .	154
6.9	An example on the incorrect alignment due to the presence of unspecified NUM feature (i.e. NUM=nil). Since NUM feature of <i>here</i> is unspecified, it can align with <i>any arbitrary number of pen gestures</i> without penalty. . . . .	154
7.1	Statistics of lexical and multimodal terms in the training set. .	158
7.2	Task goal inference accuracy using vector-space model based on different weight methods. Please note that the test set lacks expressions in task goal CHOICE OF VEHICLE (i.e. only contains 8 task goal vectors). . . . .	162
7.3	Lexical and multimodal terms with the highest LSM weights for each task goal. Terms with an asterisk (i.e. *) are the identified key terms. . . . .	165
8.1	Statistics of the lexical and multimodal terms (count by type).	188
8.2	Task goal inference accuracy using the vector-space model approach based on different weight methods with spoken terms regularization. . . . .	189
8.3	Task goal inference accuracy before and after applying spoken terms regularization. . . . .	192

8.4	Examples of the inquiry that belong to the latent semantic categories 13 and 19 in the training set for task goal BUS INFORMATION. . . . .	196
8.5	Examples of the inquiry that belong to the latent semantic categories 7, 9, 11, 12, 29 and 46 in the training set for task goal OPENING HOURS. . . . .	200
8.6	An excerpt of the term-inquiry matrix $G$ corresponding to two multimodal inputs. The weights (shown up to 2 decimal places) are obtained using Equation 8.1. Translations are italicized. .	205
8.7	An excerpt of the reconstructed term-inquiry matrix $\hat{G}$ corresponding to two multimodal inputs as in Table 8.6. The estimated weights (shown up to 2 decimal places) of $\hat{G}$ are obtained using Equation 8.4 with $R' = 263$ . Translations are italicized. .	206
8.8	Examples of inquiries that belong to the task goal of ROUTE FINDING but are incorrectly infer as TIME CONSTRAINT. Translations are italicized. . . . .	209
D.1	An example showing the normalized cost of each recognized SLR based on Equation 6.1 for the $N$ -best ( $N = 100$ ) recognition hypotheses. . . . .	253

E.1	An example illustrating the hypothesis rescoring process of based on the $N$ -best speech recognition hypotheses ( $N = 100$ ) listed in Table D.1. The first and the second SLRs <i>here</i> , should have the numeric feature NUM=nil, which can be aligned with any number of pen gesture. The second SLR, <i>these places</i> , should have the numeric feature NUM=plural, which can be aligned with more than one pen gestures. All the five pen gestures incur no cost because their coordinates coincide with the respect icons/labels. Each candidate for cross-modality integration is rescored and then the updated rank is shown for each candidate. The 98 <sup>th</sup> hypothesis pair ranked top after rescoring. . . . .	271
F.1	A significant test on the cross-modality hypotheses rescoring of the first best recognized speech hypotheses and $M$ -best pen recognition outputs ( $M = 4$ ) from the training set. . . . .	274
F.2	A significant test on the cross-modality hypotheses rescoring of the first best recognized speech hypotheses and $M$ -best pen recognition outputs ( $M = 4$ ) from the test set. . . . .	276
F.3	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the training set. . . . .	278
F.4	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the test set. . . . .	280
F.5	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and $M$ -best pen recognition outputs ( $M = 4$ ) from the training set.	282

F.6	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and $M$ -best pen recognition outputs ( $M = 4$ ) from the test set. . .	284
F.7	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of speech recognition errors. . . . .	286
F.8	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of pen recognition errors. . . . .	288
F.9	A significant test on the cross-modality hypotheses rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) and $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of both speech and pen recognition errors. . . . .	291
G.1	A list of abbreviations used in this thesis. . . . .	293

# Chapter 1

## Introduction

Human-human communication is multimodal, where people can simultaneously combine multiple modalities, including vision, hearing, speech, eye-gaze, facial expression, gestures, posture, etc., so as to deliver their message effectively and efficiently. The advantages can be achieved by the complementary and redundant relationships across modalities. Figure 1.1 provides an illustration.

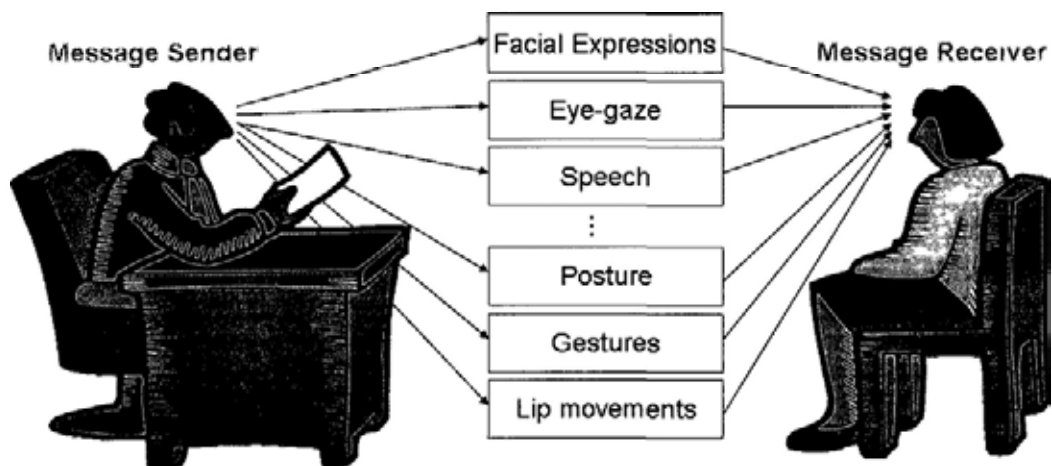


Figure 1.1: An illustration of human-human communication with multiple modalities.

Multimodal message delivery requires the message sender to divide a message into different modalities. Therefore, each modality alone carry incomplete semantics. The *complementary* semantics across modalities need to be combined in order to get the holistic interpretation of the multimodal message. For example, when a person is narrating a story about a woman who takes an umbrella as a weapon, swings it in the air and chases after a man, the person may just say “*she chases him out*” with hand movements that appear to swing an object through the air. In order to have a complete insight of the person’s thinking, we have to combine the information conveyed by both speech (i.e. the action of chasing) and hand movements (i.e. show the swinging of the weapon). Alternatively, the person may present the same thinking unimodally with speech, as: “*She takes an umbrella as a weapon and chases him out. She swings the umbrella in the air while she is chasing him.*” Figure 1.2 illustrates the scene where the person is narrating the story with complementary speech and hand movements. This example also indicates that the use of complementary modalities can simplify the spoken expression and thus enhance efficiency in human-human communication.

On the other hand, the message sender can repeat the same piece of information in different modalities. In other words, the information in multiple modalities are *redundant*. For example, when a person is describing a story, in which the man bends a tree branch backwards to the ground, the person may say “*he bends it way back*” with arm movements that appear to grip something and pull it back. The arm movements are repeating the semantic content of speech. Although the information conveyed in different modalities is mutually redundant, the repeated information can enhance the correctness of information transmission through mutual reinforcement. Figure 1.3 shows an illustration of the scene where the person is describing the story with redundant speech and arm movements.



Figure 1.2: An illustration of complementarity between speech and hand movements. In this example, the person is narrating a story. He says, “*she chases him out*” with hand movements that appear to swing an object through the air. The hand movements show the swinging of a weapon while speech conveys the action of chasing. This example is borrowed from [1].



Figure 1.3: An illustration of redundancy between speech and arm movements. In this example, the person says, “*he bends it way back*” with arm movements that appear to grip something and pull it back. The arm movements exhibit the same semantic content as presented in speech. This example is borrowed from [1].

Conventionally, human-computer interaction is unimodal. The WIMP (windows, icon, menu and pointing device) interfaces allow people to input through the keyboard, joystick, mouse, etc. Information flows in and out through one modality at a time as a stream of input/output events (i.e. the interface maintains its singularity and can only process the input/output event one by one). This is in contrast to natural human-human communication where information can be input/output through multiple modalities and processed in parallel (i.e. the interface can process multiple input/output events through different modalities at the same time).

The growing penetration of mobile devices, like the Apple® iPhone<sup>1</sup> [2], Nexus One<sup>TM</sup> <sup>2</sup> [3], iPAQ smartphone [4], etc. allows us to go beyond the desktop computers. Pervasive computing presents new requirements for human-computer interaction where computers and their screen-sizes are much smaller. There are often constraints due to usage environments (e.g. speaking in a noisy office, screen glare, etc.), constraints due to user skills (e.g. for small children, difficulties in Chinese input, etc.), and constraints due to users' physical abilities (e.g. language barriers, visually impairments, etc.). Multimodal interactions are increasingly appreciated, where users can use multiple modalities individually or in concert to overcome the constraints. For example, pen gestures make it easier for conveying spatial information, while speech communication is preferred in hands-busy, eyes-busy environment.

---

<sup>1</sup>Until the first quarter of 2010, Apple sells 42,487,000 iPhones in 11 quarters (information source <http://moconews.net/article/419-apple-sells-record-8-74-million-iphones-during-holidays/>)

<sup>2</sup>Until mid-March 2010, Google sells 135,000 Nexus One<sup>TM</sup> (information source Flurry statistics <http://blog.flurry.com/bid/31410/Day-74-Sales-Apple-iPhone-vs-Google-Nexus-One-vs-Motorola-Droid>)



## 1.1 Challenges

Multimodal user interface offers expressive power to users, but we need technologies to decode the user's intention. This presents the research problem of automatic semantic interpretation of multimodal user inputs. In other words, we wish to ask the question: *How we can recognize, integrate and interpret input from several modalities and generate a single semantic interpretation?*

### 1.1.1 Recognition of User Input in Different Modalities

There are many different combinations of modalities. For example, if we focus momentarily on speech and gestures, we can have speech and pointing gestures [5], speech and lip movements [6] [7], speech and pen gestures [8] [9], speech and mouse clicks [10] [11], speech, pen gestures, facial expressions [12] [13], etc. Generally, multimodal systems include a recognizer for each modality. The challenge we face in recognition of the user input in each different modalities is the performance, since it varies in different conditions. For example, the performance of a speech recognizer is affected by the type of microphone, usage environments, accents, speaking style, the speaker's voice, etc. We can use a directional microphone with noise cancellation function for applications in a desktop PC to ensure a better speech input quality. However, in a mobile setting, we have to use the built-in microphone, which usually has poor quality. The presence of environmental noise (especially non-stationary noise includes a person talking) also affects the speech recognition performance. Accents often decrease the performance of a speech recognizer which is trained on standard pronunciations where speaker adaptation may be required to improve the performance. Speaker independent recognition performs worse than speaker-dependent recognition. This is because different persons may differ in accents, physical properties of vocal tract, spectral characteristics for the same speech sound, etc.

Another example is based on pen gesture recognition. The challenges include segmentation and classification. We can use a range of input devices including the touch screen, tablet, mouse, etc. to capture of a trace of subsequent coordinates as pen inputs. However, how can we segment the coordinates into one or multiple pen gestures? After segmentation, how can we classify the pen gesture into which pen gesture type? The pen gesture recognition performance varies with the writing style, number of writers, number of pen gesture types, etc. Different persons may draw a symbol or write a character in different ways – e.g. an open-mouth circle versus the character “U”, a point versus a short stroke, etc. that make the pen gesture become difficult to be recognized. Similar to speech recognition, writer-dependent pen gesture recognition performance is often better than writer-independent recognition.

### 1.1.2 Message Fission

People deliver their message through different modalities in various ways. They may divide up their message across modalities in a complementary or redundant manner. A person may prefer to use speech to present his intention and to use pen to indicate the visible objects of interest. Moreover, the temporal patterns across the modalities may be different, e.g., simultaneous and sequential temporal patterns. Alternatively, a person may speak out the entire request (including both his intention and related objects) and use pen gestures to indicate the visible objects of interest on the system interface again. For example, a person may say, “What are the opening hours of this place?” to present his intention (i.e. to ask for the opening hours of a location) and then use the pen to circle a restaurant on the map on screen. This will exhibit a sequential temporal pattern. Alternatively, the person may indicate the location of interest in both the speech and pen modalities, e.g., “What are the opening hours of the Glory Restaurant?” and encircling the icon of the Glory Restau-

rant at the same time. This will exhibit a simultaneous temporal pattern. It will be a challenge to learn how people divide up their messages.

### 1.1.3 Cross-modality Integration

Another challenge in the design of multimodal systems involves learning how to integrate information from different modalities so as to obtain the user's original intention (i.e. cross-modality integration). Since people divide up their messages differently, a multimodal system should be able to take advantage of the complementary inputs and see how the information from different modalities can compensate for each other. On the other hand, a multimodal system should also be able to take advantage of redundant inputs such that the information from different modalities can reinforce with each other to ensure correct integration. However, information from different recognizers may contain errors. Therefore, a multimodal system must be robust to recognition errors.

### 1.1.4 High-dimensionality

Multimodal system offers expressive power to people to make an input into the computer. The high dimensionality of input features (e.g., the lexicon size in the speech recognizer and number of pen gesture types supported in the pen gesture recognizer) and freedom in input styles (e.g., the ways that people divide up their messages across modalities) may affect the performance (in both efficiency and accuracy) of integration and interpretation. Large amounts of training data are needed to cover all possible variations. An efficient dimensional reduction method will be needed to enhance efficient computation and reduce data requirements.

## 1.2 Thesis Goals

We address the research problems mentioned above, in the context of multimodal inputs with speech and pen gestures. These two modalities are gaining ubiquity in our daily lives, e.g. hand-held devices with global positioning systems (GPS), use of Google Maps [14], and use of speech and pen gestures to control a mobile device and indicate spatial information. Moreover, coordinated use of both speech and pen gestures enhances expressive power, especially in the communication of complex semantics in succinct form [15]. Consider the unimodal spoken inquiry:

*What is the name of the street that is five blocks south of the Yonghegong and lies to the east of the China National Museum of Fine Arts?*

may be paraphrased multimodally with substantial simplification, to become:

*What street is this? <draw a stroke on the map>*

Since speech and pen gestures are less temporally coupled, we apply semantic-level integration to process multimodal speech and pen inputs. The semantics of a multimodal input may be imprecise (e.g. a pen stroke on a map may denote a street or demarcation), incomplete (e.g. use of anaphora in “*how about the previous one?*”) or erroneous due to mis-recognitions (e.g. speech or pen gestures recognition errors). These problems motivate us to investigate the following research problems:

- characterization and extraction of features from each modality – specifically, we focus on speech and pen gestures;
- recognition of input events from each modality – specifically, spoken locative references in speech and pen gestures in pen input;
- interpretation of recognition outputs of each input event (i.e. spoken locative references and pen gestures) as their partial semantics;

- integration of the partial semantics across modalities;
- maintaining robustness against imperfectly captured inputs and misrecognitions by leveraging the mutual reinforcement and mutual disambiguation across modalities [16] [17]; and
- interpretation of the user's intention by integration across multiple modalities.

### 1.3 Thesis Organization

This thesis begins with some background information about multimodal systems and a brief mention of related studies in cross-modality semantic interpretation and integration. Chapter 2 introduces the variety of modalities in multimodal system, relationships across modalities, related study on the semantic interpretation/integration methods and evaluation methods of multimodal systems. In order to support our investigations, we have designed and collected a multimodal corpus for navigational inquiries. Our work in the corpus design and collection is presented in Chapter 3. Chapter 4 describes our findings in an exploratory data analysis of the collected multimodal corpus, including the characterization, representation and relationships between modalities. Details related to the partial interpretation of input events from each modality are also presented in this chapter. The proposed cross-modality semantic integration framework is introduced in Chapter 5, where we applied the framework to perfect and erroneous recognition outputs so as to obtain upper and lower bounds of the semantic integration performance. The perfect recognition outputs are referring to the manual transcriptions while the erroneous recognition outputs are the recognition outputs automatically generated by speech and pen gesture recognizers. We extended the cross-modality semantic integration framework with hypothesis rescoring to gain robustness

against imperfectly captured inputs and recognition errors, and the details will be presented in Chapter 6. Chapters 7 and 8 address our work in developing a semantic analysis framework for task goal inference. Finally, this work is concluded in Chapter 9, and we will also mention some possible future work.

## Chapter 2

# Related Work

In Chapter 1, we have stated our motivation and goals in recognition of speech and pen gestures, partial semantic interpretation and cross-modality semantic integration. This chapter presents related work in multimodal user interfaces and cross-modality semantic integration. We would like to start by exploring the variety of modalities with an focus on touch/pen-based modality and visual-based modality. Previous work in relationship across modalities will also be presented. One of the goals of this thesis is to develop a cross-modality semantic integration framework, so previous work in the multimodal fusion (i.e. semantic interpretation/integration methods) will also be described. Finally, we will review some work on the evaluation of multimodal user interface.

### 2.1 Variety of Modalities

Since the appearance of the “Put-that-there” [5] system, which processed speech in parallel with manual pointing during object manipulation, much research effort has been devoted to the development of multimodal user interfaces with various combinations of modalities such as speech and lip movement, speech and eye-gaze, speech and head movement, etc. We focus on two main categories of multimodal user interfaces: pen/touch-based modalities

and visual-based modalities.

The category of speech with pen/touch-based modalities for interactions with graphical, image and video data usually involves a pointing device or a finger/pen on a touch-sensitive screen [18]. This category supports visual-spatial applications involving map-based interactions [19], sketching applications [20], character [8] and handwriting [21]. Some example systems include Quick-Set [17], which runs on a handheld PC for military simulation and medical informatics that enable the user to create and position entities on a map through speech and pen (including drawn graphics, symbols and pointing) gestures; RealHunter<sup>TM</sup> [22] for real-estate information, which helps users to find residential properties through speech and pen gestures (including highlight, pointing and circling, etc.); city navigation systems such as Voyager [23] [24], which provides navigation assistance and traffic information for Boston; and MATCH (Multimodal Access To City Help) [25], which provides navigation for restaurants, points of interest and subway information for New York and Washington, DC. Other applications include HCWP (Human-Centric Word Processor) for voice dictation of radiology reports [26]; WITAS (Wallenberg laboratory for research on Information Technology and Autonomous Systems) [10] [11] for communicating with unmanned aerial vehicles [27] using speech and mouse clicks; MiPad (Multimodal Interactive Personal Assistance Device) [28] for personal information assistance using speech and pen gestures; Miki [29] for simultaneous recognition and understanding to solve a mathematical problem through integration of speech and fingertip movements; COMIC (CONversational Multimodal Interaction with Computers) [30] [31] for the applications of architectural design through the use of speech, writing and drawing.

The category of speech with visual-based modalities includes lip reading, facial expressions, eye-gaze and three-dimensional (3D) gestures. They are usually perceived by computer vision technologies. Speech and lip reading



can be applied in automatic audio-visual speech recognition [7] so as to sustain recognition performance especially in noisy ambient conditions. Speech (prosody) and facial expressions (which is influenced by both an affective state and speech content) are fused in audio-visual affect recognition [32] for tracking of the user's affective states in human-computer interaction. Eye-gaze may indicate a deictic reference and/or the focus in conversation, tracking of eye-gaze behavior can check whether a user is engaged in the conversation [33]. Included in 3D gestures are the head, hands, fingers and more generic body movements. Speech (voice-print), face and fingerprint can be regarded as digital personal identity. Hui [34] combines these three modalities with a dynamic weighting scheme using fuzzy logic for speaker verification. Head movements may directly convey a message, e.g., signifying agreement by nodding, and are used extensively in face-to-face communication. Head movement recognition performance can be improved by integrating the predictions that are made based on the modalities of speech (i.e. lexical and punctuation features) and head movements (i.e. output of head gesture recognizer) [35]. Hand and body movements are suitable modalities for virtual-reality applications [36]. In addition, there are multimodal applications in meeting recordings, such as the AMI project [37] that uses the modalities of speech, eye-gaze, head and hand gestures, body movement, facial expressions, etc., allowing users to find information they are interested in quick from a recorded meeting [38] in a smart room. SmartKom [12] [13] is a large-scale multimodal dialogue system that combines speech, pen gestures and facial expressions in interfaces for mobile computers, public information kiosks and smart homes. User can interact with the system through a combination of speech, gestures and facial expressions.

## 2.2 Relationships across Modalities

This section discusses the properties that we need to consider for the development of multimodal user interface. Development of the multimodal interface depends on the knowledge on the features of natural human communication methods, cognitive status of people that affects their choice of modalities, the natural integration patterns that people use to combine different modalities and the usability of multimodal interactions that influence the design of a multimodal user interface.

### 2.2.1 Speech-Gesture Correlation

During human-human communication, people often use hand movement in parallel with speech so that speech and hand gestures are complementary with each other [1]. Concurrent hand movements can be classified into four types: iconic gestures present images of concrete objects and/or actions including size, shape, trajectory, direction, etc.; metaphoric gestures present images of the abstract of ideas as form and/or space; deictic gestures are usually related to pointing that entails locating entities and actions in space to a reference point; and beats where the hand moves along with the rhythmical pulsation of speech, which can be used to signal something important. However, a gesture may belong to more than one type.

Chen [39] focused on the iconic and deictic gestures and analyzed the correlation between speech and hand gestures on prosodic and lexical levels for multimodal input fusion and gesture classification. The study showed that about 65% of the deictic gestures are synchronized in time with the peaks of the delta pitch contours of speech, and a deictic gesture is likely to occur given a peak in the delta peak of speech. It also showed that following the lexical pattern allows them to predict an upcoming deictic gesture at about 75% confidence. The prosodic and lexical features found in this work can be

incorporated into the integration and fusion mechanisms of speech and gestures.

### 2.2.2 Cognitive Status and Form of References

During communication, humans often refer to something using references (or referring expressions). These references may be ambiguous or incomplete. People may be able to understand each other if the message receiver knows the referent's *cognitive status* so as to identify the intended referent. Previous research efforts proposing six cognitive statuses that are relevant to the form of references in Givenness Hierarchy [40] are shown in Table 2.1, together with their characterizations:

	in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
English	<i>it</i>		<i>that</i>		<i>that</i> N		<i>the</i> N		indefinite		<i>a</i> N
			<i>this</i>						<i>this</i> N		
			<i>this</i> N								
Chinese	∅		這( <i>this</i> )				那 N				— N ( <i>a</i> N)
	他/她/它 ( <i>he,</i> <i>she, it</i> )		那( <i>that</i> )				( <i>that</i> N)				∅ N
			這 N								
			( <i>this</i> N)								

Table 2.1: Correlation between linguistic form and status in Givenness Hierarchy where N stands for noun and ∅ means zero. ∅ N (in the last column) indicates the use of a noun only and no article is required in Chinese (e.g. *dog*). The arrows indicate that the statuses are ordered from most restrictive to least restrictive with respect to the possible referents they include.

**Type identifiable:** the message receiver is able to access the representation of the type of the object described by the reference.

**Referential:** the message deliverer (i.e. source) intends to refer to a particular object(s).

**Uniquely identifiable:** the message receiver can identify the deliverer's intended referent on the basis of the nominal alone.

**Familiar:** the message receiver is able to uniquely identify the intended referent because he has a representation of it in memory.

**Activated:** the referent is represented in current short-term memory.

**In focus:** the referent is not only in short-term memory but also at the current center of attention.

Kehler [41] applied the first four statuses of Givenness Hierarchy (i.e. in focus, activated, familiar and uniquely identifiable) to multimodal human-computer interaction of travel guide application. Subjects were asked to plan activities and plan places to stay, see, and dine using speech and pen gestures for a hypothetical trip to Toronto. This work found that a simple decision list procedure can be used for reference resolution as shown below:

- If an object is gestured to, choose that object.
- Otherwise, if the currently selected object meets all semantic type constraints imposed by the referring expressions, choose that object.
- Otherwise, if there is a visible object that is semantically compatible, then choose that object.
- Otherwise, a full noun phrase was used that uniquely identified the referent.

He found that subjects inferred their thoughts only from the information on the visual display, which marked only the cognitive statuses of in focus (i.e. selected) and activated (i.e. unselected but visible). Subjects only distinguish

between unselected referents by either a multimodal expression (i.e. referential expression together with a disambiguating gesture) or a full and uniquely-identifiable definite unimodal (i.e. speech-only) expression. It also found that speech-only human-computer communication will result in less efficient reference than multimodal communication.

### 2.2.3 Integration and Synchronization of Input Modalities

Besides the relationship between cognitive status and references, multimodal interactions were examined in [17] and [42] on the basis of user preference, task action, linguistic content and integration patterns.

Analysis shows that subjects have a strong preference to interact multimodally with map-based systems. All of them used at least once multimodal input of speech and pen gestures during a task. Subjects tend to use multimodal interactions so as to reduce their cognitive load when tackling tasks with increasing difficulty and communicative complexity.

Subjects tend to use multimodal inputs for spatial location commands (including adding object(s), moving object to a new location, modifying specific routes or spatial areas and calculating distance between two locations), which requires spatial location information (86% of the multimodal inputs as mentioned in [17]).

The majority (98%) of the multimodal construction conformed to the typical subject-verb-object order of English. It shows that the main difference between multimodal and unimodal (i.e. speech only) input is the position of the locative descriptor, where the locative descriptor always at the beginning of a multimodal input but at the end of a unimodal input.

The majority (86%) of the multimodal constructions show a “draw and speak” pattern. It shows that the pen always precedes speech in both simultaneous and sequential inputs. Moreover, the maximum lag between speech

and a pen gesture is less than three seconds 97% of the time.

#### 2.2.4 CARE Properties

The CARE properties proposed in [43] are used to characterize and assess aspects of multimodal human-computer interactions. CARE stands for complementary, assignment, redundancy and equivalence. They are the properties that influence the design and implementation of multimodal user interface. A description of each property is as follows.

- (1) *Complementary* use of two modalities can generate a complete inquiry with the necessary information. For example,

S: 從“這個地方”到“這四個大學”要多久

P:           • (a point)           ○ (a circle)

Translation: *How much time will it take from “this location” to “these four universities”?*

We only know that the user wants to go from one location to four universities from the spoken inquiry. We also get the name of five locations from pen gestures. Therefore, we need to integrate information from both modalities to be one.

- (2) *Redundancy* can ensure correct semantic interpretation of the locations as in the example:

S: 從“人大”到“北郵”怎麼走最快

P:           •                   •

*What is that shortest route from the “Renmin University of China” to “Beijing University of Posts and Telecommunications”?*

Since locations obtain from both modalities should be the same in this case, we can ensure that the locations interpreted are correct.

- (3) *Assignment* property is applied on the communication goals. In this work,

speech is the dominant modality because speech indicates the status of the interaction (i.e. task goal and dialogue act for understanding). Therefore, the speech modality is assigned to present the type of information that the user is interested in.

- (4) The two modalities are *equivalent* on the expression of location (spatial information). The user can either speak out the location name or point on it during interaction. The process of joint interpretation/integration should also incorporate the processes of mutual reinforcements and mutual disambiguation across modalities [17] due to their complementarity and redundancy.

U-CARE properties are a counterpart of the CARE properties where U-CARE properties are the CARE properties of the user. U-CARE properties are concerned with the user's choice between different modalities. Usability of a multimodal user interface can be evaluated by considering its compatibility with U-CARE properties as mentioned below:

U-complementarity means user provides part of the information in one modality and the remaining one or more further modalities. The compatible condition is that system-complementarity and U-complementarity modalities are the same.

U-redundancy means all modalities available to the user are used. The compatible condition is the system-assignment modality is among the U-redundancy modalities. Moreover, there is at least one common modality between system equivalent modalities and U-redundancy modalities.

U-assignment means that user requires a particular modality. The compatible condition is that the U-assignment modality be among system-redundancy modalities.

U-equivalence means that user is prepared to use any one of the modalities. The compatible condition is that the system-assignment modality be among

the U-equivalence modalities of the user. Moreover, there is at least one common modality between system equivalent modalities and U-equivalence modalities.

### 2.3 Semantic Interpretation/Integration Methods

Multimodal integration is the technology that integrates information or semantics from two or more human communication modalities. The integration gives rise to an interpretable output with holistic semantics. There are mainly two approaches on the integration of multiple modalities: feature-level integration and semantic-level integration. Information processing of a multimodal input starts with *within*-modality processing (i.e. recognize the input event from each modality and interpret the recognition output so as to generate partial semantic interpretation for each modality). Then, we can perform *cross*-modality processing (i.e. integrate jointly the partial semantics from different modalities) to generate a holistic interpretation. This is referred as semantic-level integration (see Figure 2.1). Alternatively, we can adopt feature-level integration, where recognition outputs across modalities are first integrated and then interpreted. Feature-level integration is often applied early so as to combine highly dependent and synchronized input modalities, e.g., speech and lip movements while semantic-level integration is performed at the word or phrase level of a multimodal expression.

Feature-level is appropriate for highly dependent and closely temporally synchronized input modalities, e.g. speech and lip movements. The features from one modality influences the recognition of features in the other. The two modalities are usually combined using histogram techniques, multivariate Gaussians, artificial neural networks or hidden Markov models [44]. Significant improvement in robust speech recognition performance using both speech and visual information (lip movement) was showed in [6]. The problem of bimodal



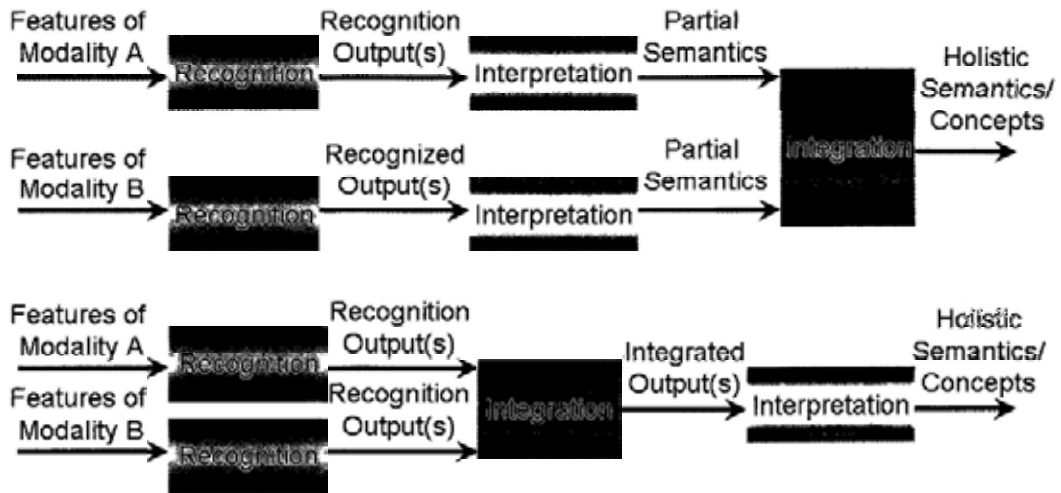


Figure 2.1: Semantic-level integration (top) vs. feature-level integration (bottom).

(speech and pen gestures) character auto-completion (CAC) has also involved feature fusion. Significant improvement can be obtained by combining handwriting CAC and speech recognition candidates in the posterior sense [8].

Semantic-level integration is more suitable for modalities that are less coupled temporally and is performed at the word or phrase level of a multimodal expression. It is used to integrate partial semantic information (hypotheses) from each modality together with some contextual information to be a reasonable interpretation of the user’s multimodal input. However, it can be decomposed into two sub-problems: *representing* partial semantic information from each modality and *integrating* pieces of partial semantics with other contextual information into a holistic interpretation of the user’s intention. The “Put-that-there” system [5] and QuickSet [9] are two examples of a semantic-level integration system.

Previous approaches toward cross-modality semantic integration / interpretation of multimodal input include frame-based heuristic integration, unification parsing, hybrid symbolic-statistical approach, weighted finite-state transducers, probabilistic graph matching and the salience-driven approach. We

will describe them in greater detail below.

Frame-based heuristic integration [45] [23] uses an attribute-value data structure to represent partial semantics from each input modality and each type of contextual information. The data structures are then merged according to top-level control heuristics and pattern matching techniques that incorporate temporal difference and contextual information. Research work in [45] devised the “melting pot” representation that encapsulates types of structural parts of a multimodal event with a three-step procedure handling simultaneous input by microtemporal fusion, sequential input by macrotemporal fusion and context-based fusion by contextual fusion. Wang [23] developed a multimodal context-resolution module for resolving anaphoric and deictic references based on syntax and semantics in spoken language. However, Wang does not claim to support events with multiple gesture-based selections.

Unification parsing was proposed in [46] [47]. This approach represents  $N$ -best speech/pen recognition hypotheses as typed feature structures. Temporally compatible multimodal combinations are combined semantically by multi-dimensional chart parsing using a declarative unification-based grammar. The complex grammar rules are written by hand and encapsulate semantics from both modalities, as well as a set of spatial and temporal constraints for multimodal integration. Authoring rules require a high level of expertise.

The hybrid symbolic-statistical approach was proposed in [48] for Quick-Set. This approach aims to statistically refine unification-based parsing with probabilities and confidence scoring of the features structures in order to account for co-relations between modalities. This approach filters for semantically plausible associations across modalities, followed by weighted interpolation of the probabilities from the individual feature structure. Weights are trained by the Members-Teams-Committees (MTC) technique [49]. Since multiple sets of trained weighting parameters (i.e. parameters at each level of

the MTC hierarchy) are used in the MTC technique, much training data is needed. SmartKom [12] also applies a unification approach on the recognition hypotheses graphs for each modality with adaptive confidence scoring [13].

Weighted finite-state transducers (FSTs) was proposed in [25] [50] for the MATCH system. This approach encodes syntactic and semantic information to offer tight coupling across modalities, with FST weights as trained from data. This approach also requires the development of a multimodal grammar used with a FST. The grammar has non-terminals that are atomic symbols and terminals are three-tuples consisting of spoken words, gestures and their combined meaning. The grammar contains many rules and is relatively complex. Again, authoring such a grammar requires specialized skills. Furthermore, the work in [50] indicated that the approach has difficulty in handling general plural expressions, which may be integrated with a multitude of possible sequences of gestures.

Probabilistic graph matching was proposed in [51] [22]. This approach incorporates semantic, temporal and contextual constraints to combine information from multiple input modalities, where the information is represented as attribute relational graphs (ARGs). Each graph node encodes semantic/temporal information and each edge encodes semantics/temporal relations. Integration includes maximizing the node match probabilities between ARG from speech and the ARG from pen input. The work in [22] indicated that a higher number of referring expressions may cause the approach to become intractable because the graphs increase in size.

The salience-driven approach was proposed in [52]. This is an  $n$ -gram language model that incorporates a salience distribution based on the pen gestures to constrain the bigram probability for understanding spoken language. Trained weights are used in a probabilistic context-free grammar (PCFG), which is applied in language modeling. The large number of weights to be

trained demands much training data.

## 2.4 Evaluation Methods

There are many components in a multimodal system. Hence, evaluation of a multimodal user interface may be carried out at the component-level or system-level.

### 2.4.1 Component-level Evaluation

Evaluation of the system components can re-use the evaluation methods used in various sub-fields. We can evaluate the system components based on recognition accuracy or error rate, for example, evaluation of speech recognition [53], evaluation of pen gesture recognition and evaluation of handwriting recognition. We can also evaluate a component based on user perception, for example, of talking head [54] and text-to-speech synthesizer.

### 2.4.2 System-level Evaluation

As mentioned in Section 2.3, there are mainly two approaches in the integration of multimodal inputs: feature-level integration and semantic-level integration. System-level evaluation can be further divided into two categories according to the integration approach.

Evaluation of multimodal user interface with feature-level integration is similar to the evaluation of system components. Evaluation criteria including recognition accuracy, error rate, false rejection rate and user perception, etc. can be used. For example, the audio-visual automatic speech recognition system developed in [7], which combines speech and lip movement for robust speech recognition, uses word error rate as the performance evaluation criterion.

An appropriate evaluation metric has to be defined for the evaluation of multimodal user interface with semantic-level integration according to the nature of the system and availability of a test set. Some possible metrics include task completion time, task success rate, number of turns, naturalness, user satisfaction, cost, etc. These evaluation metrics can be obtained through different evaluation approaches, including user-based, theory-based and expert-based evaluations [55].

- **User-based Evaluation** Benchmark evaluation, simulation studies and user studies are examples of user-based evaluation. Benchmark evaluation requires the collection of a test set for performance evaluation. It is suitable to test the overall performance of a system based on a set of multimodal inputs. Work in [56] used the benchmark evaluation method for performance evaluation. Simulation studies can simulate a multimodal system before implementation of a working system. The Wizard-of-Oz technique has been widely used for simulation study [17]. User study requires a multimodal system prototype for evaluation, but the user inputs collected during user study can build up a multimodal database for benchmark evaluation afterwards. For example, MiPad [57] performed a user study and used task-completion time and user satisfaction as evaluation criteria to study whether their *Tap and Talk* interface can add value to the PDA user interface.
- **Theory-based Evaluation** The predictive model is an example of theory-based evaluation. It predicts user behavior or performance variables based on pre-defined assumptions and model parameters. It allows evaluation of the multimodal system at the design stage so as to improve the design before implementation.
- **Expert-based Evaluation** This type of evaluation requires a human expert to evaluate whether the system matches with the pre-defined de-

sign criteria or established design heuristics in a structured way using a prototype system.

## 2.5 Chapter Summary

This chapter presents the previous work that are related to the cross-modality semantic integration framework. The framework is motivated by the increasing need of multimodal user interfaces where users can use multimodal modalities individually or in combination to overcome the constraints due to usage environment, users and user's skills. We also explore the variety of modalities with a focus on the two main categories: pen/touch-based modalities and visual-based modalities. Since multimodal interfaces consist of two or more modalities, we need to consider their properties and correlation for the development of multimodal user interface. Therefore, relationships across modalities, including features of natural human communication methods, cognitive status of people that their choice of modalities and the natural integration patterns that people use to combine different modalities and the characterizations of multimodal human-computer interactions (i.e. the CARE properties) are discussed in this chapter. One of the goals of this thesis is to develop a cross-modality semantic integration framework, so previous work in the semantic interpretation/integration methods are described. The proposed framework contains many components so details of the component-level and system-level evaluations are also discussed in this chapter.

## Chapter 3

# Multimodal Corpus

This chapter describes our work in the design and collection of a multimodal corpus of navigational inquiries. The multimodal corpus is a collection of bi-modal user inputs that has been organized, transcribed and annotated to support our investigation. The design principles, data collection procedure, corpus statistics and annotation methods will be presented in this chapter.

### 3.1 Information Domain

The current investigation is cast in the information domain of navigation around Beijing. Inquiries involving locative information often induce multimodal user input. We downloaded thirty two maps from the Internet,<sup>4</sup> covering seven districts in Beijing. Figure 3.1 shows the coverage of the maps downloaded. Figure 3.2 shows an example of the map. We identified about 4,652 locations associated with icons and labels on the thirty two maps. For each icon, we annotated their positional coordinates, corresponding to the four corners of the icon. We also categorized the icons according to “location types” and “sub-types”. There are seven location types in all, including TRANSPORTATION (e.g. a bus stop), LAND AND WATER (e.g. a river), PO-

---

<sup>4</sup><http://one.5i.net.cn/html/bjmap/bjmap/bjmap.htm>

LITICAL FEATURES (e.g. a district office), LEISURE FACILITIES (e.g. a park), PUBLIC FACILITIES AND SERVICES (e.g. a hospital), SCHOOLS AND LIBRARIES (e.g. a university) and MAJOR BUILDINGS (e.g. a shopping center). Each location type is further organized into two to twelve “subtypes”. For example, the location type TRANSPORTATION contains the subtypes *road, street, train station, railway station, railroads, bus stop, bridge, intersection, highways, elevated highway, elevated road* and *road under construction*; while SCHOOLS AND LIBRARIES consists of *universities, institutes* and *libraries*. The complete list of location types and subtypes are shown in Table 3.1. For a given location type and subtype, there can be multiple instances of domain-specific data entries. For example, the location type of TRANSPORTATION and subtype of *street* will include all the street names on the map.

location type (LOC_TYPE)	subtypes
TRANSPORTATION	<i>train station, railway station, railroads, bus stop, bridge, intersection, highways, elevated highway, road under construction, elevated road, street</i> and <i>road</i>
LAND AND WATER	<i>occupied land, unoccupied land, lake, river</i> and <i>catch-water</i>
POLITICAL FEATURES	<i>capital city</i> and <i>district office</i>
LEISURE FACILITIES	<i>scenic shop, scenic spot, hotel, stadium, museum, theater, exhibition center, recreational area, green area</i> and <i>parks</i>
PUBLIC FACILITIES AND SERVICES	<i>news agency, hospital, temple</i> and <i>heritage</i>
SCHOOLS AND PUBLIC LIBRARIES	<i>university, institute</i> and <i>library</i>
MAJOR BUILDINGS	<i>shopping center, hotel</i> and <i>building</i>

Table 3.1: A complete listing of location types and their corresponding subtypes.



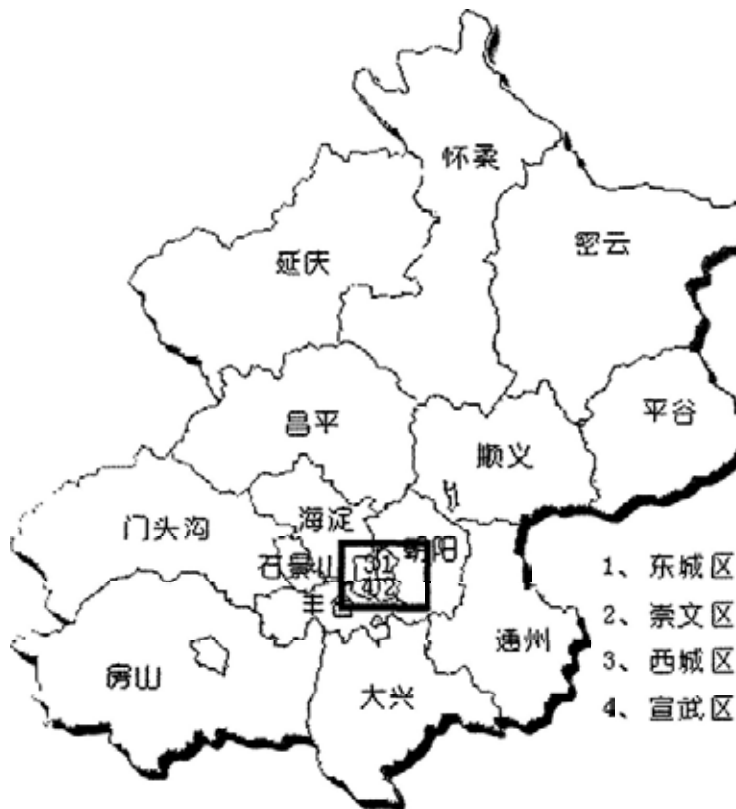


Figure 3.1: The map of the Beijing City. The coverage of the maps we downloaded from the Internet is highlighted in blue. They are Haidian District, Xicheng District, Dongcheng District, Chaoyang District, Fentai District, Xuanwu District and Chongwen District.

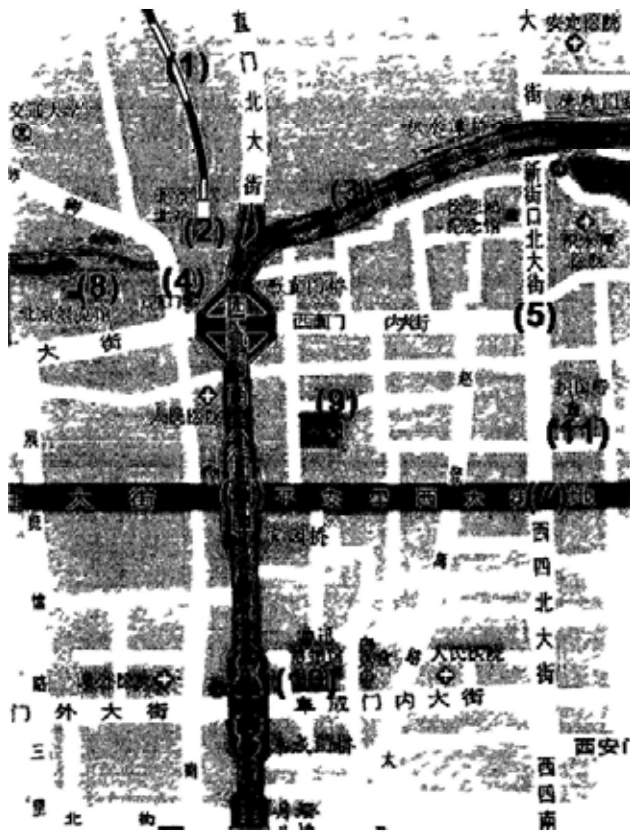


Figure 3 2: A map downloaded from the Internet. The numbers highlight some examples of location icons for the LOC\_TYPE of TRANSPORTATION, including subtypes of (1) *railroad*, (2) *train station*, (3) *elevated highway*, (4) *railway station*, (5) *intersection*, (6) *bridge* and (7) *highway*; location icons for LOC\_TYPE of LEISURE FACILITIES, including subtypes of (8) *exhibition center*, (9) *green area* and (10) *museum*; and location icon of PUBLIC FACILITIES AND SERVICES, including (11) *temple*

We also conducted a quick survey involving ten people regarding typical inquiries from users who are trying to navigate around Beijing. The survey can be found in Appendix A. These inquiries generally target nine information goals, including:

- BUS INFORMATION
- CHOICE OF VEHICLE
- MAP COMMANDS
- OPENING HOURS
- RAILWAY INFORMATION
- ROUTE FINDING
- TIME CONSTRAINT
- TRANSPORTATION COSTS
- TRAVEL TIME

Based on these information goals, we designed specific tasks (32 tasks covering 7 location types) such that each induces a subject to compose multimodal inquiries. The tasks cover various numbers of locations (which increases from zero to six locations) and different combination of location types (multiple locations in the same location types or multiple locations in different location types). Table 3.2 shows an example task and a multimodal input composed by a subject during data collection.

## 3.2 Data Collection Procedures

We invited 23 Mandarin-speaking subjects to participate in data collection. In an initial briefing session, each subject is provided with the background

Information goal: TRAVEL TIME
<p>Task: 告知系列你所在的位置，查詢從那裡順序到另外四所大學需要多長時間。</p> <p><i>Specify your current location. Find the time it takes to travel to four universities of your choice.</i></p>
<p>Multimodal input (<math>\bullet</math> denotes a point and <math>\rightarrow</math> denotes a stroke)</p> <p>S. 我在“北郵”。</p> <p><i>P</i>            <math>\bullet</math></p> <p>S. 從“這裡”出發順序到“這個大學”“這個大學”“這個大學”“這個大學”要多久？</p> <p><i>P:</i>            <math>\longrightarrow \longrightarrow \longrightarrow \longrightarrow</math></p> <p><i>I'm at "BUPT". From "here", I want to visit "this university", "this university", "this university" and "this university" in order. How long will it take?</i></p>

Table 3.2: An illustrative example for multimodal data collection with speech (*S*) and pen gestures (*P*). Translations are italicized.

information of the current work and the tasks he/she is requested to perform. Each subject is presented with an instruction sheet listing the set of 32 tasks (as shown in Appendix B). For each task, the subject is asked to formulate a multimodal input that may involve up to  $n$  locations.<sup>5</sup> The subject may refer to locations by speech (i.e. spoken locative references) and/or by pen gestures. Both speech and pen inputs are recorded directly by a Pocket PC. In some of the tasks, the Pocket PC provides a specific piece of contextual information, i.e. the current location indicated with a red cross on the map. This is illustrated by icon (1) in Figure 3.3. The subjects are also informed of several possible options:

- that a spoken locative reference may be deictic<sup>6</sup> (e.g. 這裡 *here*; 這四所大學 *these four universities*); elliptic<sup>7</sup> (e.g. 到這個公園要走多久 *how long does it take to walk to this park*) or anaphoric<sup>8</sup> (e.g. 從我的所在地到王府井要多久 *how long does it take to go from my current location to Wangfujing*) where the subject's current location can be found from contextual information;
- that a pen gesture may be a point, a circle or a stroke (with a pen-down gesture followed by a pen-up gesture).

During the briefing session, we showed the subjects a few examples of different types of spoken reference and sample usage of different pen gesture

<sup>5</sup> $n$  is constrained to a maximum value of 6.

<sup>6</sup>A "deictic phrase" is a "a key phrase specifying identity or special or temporal location from the perspective of a speaker or hearer in the context in which the communication occurs" - WordNet<sup>®</sup> [58]

<sup>7</sup>An "elliptic phrase" means "there is a omission of a word or phrase that is necessary for a complete syntactical construction but not necessary for understanding" - WordNet<sup>®</sup> [58]

<sup>8</sup>Anaphoric refers to "the use of a pronoun or similar word instead of repeating a word used earlier" - WordNet<sup>®</sup> [58] where the interpretation of an anaphora can be from the same input, contextual information or dialog history.

types. This is based on the examples listed in Table 3.3.<sup>9</sup> Then, we sent the instruction to the subjects and asked them to write down the multimodal inquiries they decided. They are also allowed to revise and re-compose their multimodal inquiries during the recording session, in order to clearly express the intended task semantics and constraints. However, they are not allowed to have discussion among themselves. This is used to avoid the subjects from copying the multimodal inquiries decided by another. Subjects are asked to indicate (based on their original intentions) the correspondences between the spoken locative references (e.g. here, the nearest station, etc.) and pen gestures after the recording session by marking on the instructions they provided - it is used as the reference for our annotation on cross-modality pairings. An example of the instruction provided by a subject is shown in Appendix C.

### 3.3 Data Collection Setup

The recording session is carried out individually for each of the 23 subjects in an open office (which has normal level of background noise). The data collection setup involves a Pocket PC with a system interface (Figure 3.3). Speech input is recorded by the built-in microphone of the Pocket PC. Pen gestures are input with a stylus. The Pocket PC interface includes several soft buttons: The *START* button should be pressed to launch the automatic system logging procedure that records the speech signal, the pen gestures and the timing information between the two modalities. The interface also contains a *STOP* button (which will only be visible after clicking the *START* button). It is only used to stop the logging procedure and save all the system log and audio file. Table 3.4 shows the logged data corresponding to the example given in

---

<sup>9</sup>The map used in Table 3.3 is borrowed from the website of DiscoverHongKong DiscoverHongKong - Touring Around - Hong Kong Walks - Yau Ma Tei and Mong Kok [http://www.discoverhongkong.com/eng/touring/hkwalks/ta\\_walk\\_lmap02.pdf](http://www.discoverhongkong.com/eng/touring/hkwalks/ta_walk_lmap02.pdf)


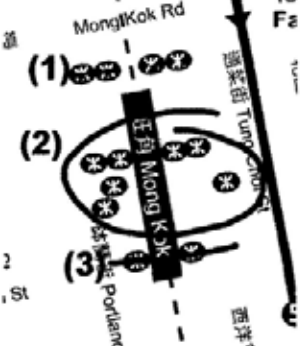
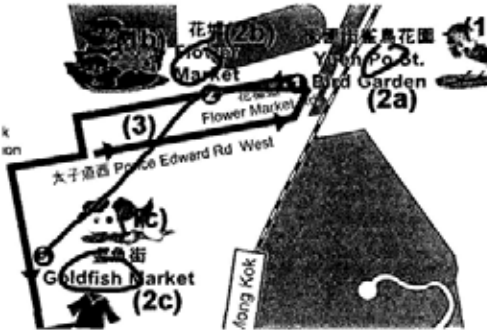
Gestures	Illustrations
<b>Select a location</b>	
(1) Point (e.g. point on the Temple St. Night Market)	
(2) Circle (e.g. circle the Tin Hau Temple)	
(3) Stroke (e.g. highlight the label of the Jade Market)	
<b>Select multiple locations</b>	
(1) Point (e.g. point on four MTR exits)	
(2) Circle (e.g. circle seven MTR exits)	
(3) Stroke (e.g. highlight two MTR exits)	
<b>Indicate a route</b>	
(1) Point (e.g. sequentially point at the icons of three locations, i.e. 1a, 1b and 1c)	
(2) Circle (e.g. circle the labels of three locations, i.e. 2a, 2b and 2b, sequentially)	
(3) Stroke (e.g. use a multi-stroke to link up the three locations)	

Table 3.3: Examples given to subjects as illustrations of the use of different pen gestures.

Table 3.2. Pressing the NEXT button displays the map of the next task.

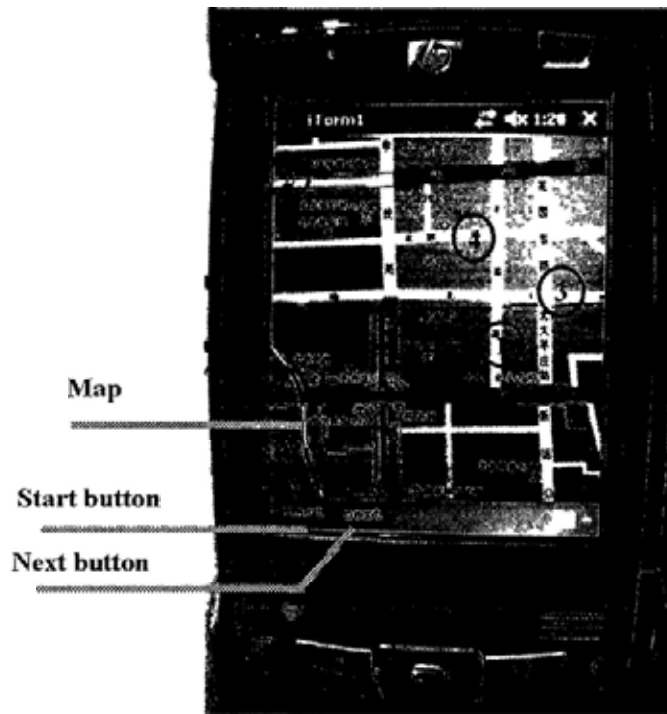


Figure 3.3: Data collection interface of the Pocket PC, augmented with soft buttons for logging functions (START/STOP) and loading the NEXT map. The numbers highlight some examples of location icons: (1) subject's current location (i.e. the red cross); (2) a university; (3) a road and (4) a hospital.

### 3.4 Corpus Statistics

We have collected 1,518 inputs from 23 subjects in all. Among these, 1,442 are multimodal and 76 are speech-only inquiries. All speech and pen data have been manually transcribed. The transcription was done by one transcriber. The process of transcription begins by transcribing of speech part of an input (i.e. listen to the audio file recorded). Then, the transcriber marks all the pen gestures and labels the multimodal pairings by looking at the multimodal



<p><b>Log for speech</b> (<i>with start and end times and the audio filename</i>)</p> <p>start: 46019 end: 46030 \Program Files\DC\AudioFile10.wav</p> <p><b>Log for pen</b> (<i>with each gesture numbered in the order of occurrence, the recognized gesture type, start and time times and x-y coordinates of the pen down and pen up actions.</i>)</p> <p>0- point start: 46022 end: 46022 from: (152,182) to: (152,182)</p> <p>1- stroke start: 46022 end: 46024 from: (152,182) to: (69,69)</p> <p>2- stroke start: 46024 end: 46025 from: (69,69) to: (70,24)</p> <p>3- stroke start: 46025 end: 46026 from: (69,24) to: (95,12)</p> <p>4- stroke start: 46026 end: 46028 from: (93,12) to: (101,61)</p>
--

Table 3.4: Example of logged data for multimodal input based on Table 3.2. Explanations are in italics.

input rendered by a home-grown visualizer.<sup>10</sup> An example of the manual transcription output is shown in Table 3.5. Utterance lengths range from 2 to 54 Chinese characters, covering a vocabulary of size 521 with domain-specific named entities and spoken locative references (SLRs). A user input may consist of between zero (i.e. speech only input) to six pen gestures. Pen gestures may be consisted of the types of point, circle or stroke. Short inputs are typically map commands (e.g. 縮小 *zoom in*). In general, long inputs include several direct locative references. Both of the longest and shortest inputs are multimodal inquiries. Details of the multimodal corpus are given in Table 3.6. We randomly divide the 23 subjects into two disjoint data sets. The training set consists of 16 subjects and has 1002 inputs (i.e. 70% of the multimodal inquiries). The test set consists of 7 subjects and has 440 inputs (i.e. 30% of

<sup>10</sup>A video of the multimodal input rendered by a home-grown visualizer is shown at <http://www.se.cuhk.edu.hk/~pyhui/visualizer.htm>

the multimodal inquiries). The training set of our corpus has 2,425 spoken locative references and 2,564 instances of pen gestures in total. Figure 3.4 shows the distribution of multimodal inquiries with different numbers of manually transcribed SLRs and pen gestures in the training set. It shows that around 74.5% (746/1002) of the multimodal inquiries in the training set have an equal number of SLR and pen gestures. Further analysis will be done on the one-to-one correspondence between the two modalities.

<b>Reference transcription:</b>	
S:	我現在在“這裡”從“這裡”到“這四個地方”可以怎麼走？
	<i>I'm now at "here". How can I go from "here" to "these four places"?</i>
P:	• • • • •
<b>Manual transcription output:</b>	
S:	我現在在 <這裡> 從 <這裡> 到 <這四個地方> 可以怎麼走
P:	<POINT> <POINT> <POINT> <POINT> <POINT>

Table 3.5: An example of manual transcription output. All the SLRs and pen gestures are indicated with <...>.

### 3.5 Manual Annotation of Cross-Modality Pairings

We have also manually annotated the cross-modality pairings between an SLR and a pen gesture for the multimodal expressions for performance analysis. These pairings are decided based on human judgment (i.e. our oracle), with the objective of obtaining a holistic and coherent semantic interpretation for the bimodal input. The cross-modality pairings annotated are considered as our oracle transcriptions because the annotator is regarded as the “human” system that we are targeted to develop. An illustration on the nature of the annotator is shown in Figure 3.5.

Number of subjects	23 subjects
Number of inquiries collected	1,518 inquiries
Number of speech-only inquiries	76 inquiries
Number of multimodal inquiries	1,442 inquiries
Number of multimodal inquiries in the training set	1002 inquiries
Number of multimodal inquiries in the test set	440 inquiries
Minimum number of characters in an inquiry e.g. 縮小 <i>zoom in</i> , 放大 <i>zoom out</i> , 往左 <i>move to the left</i>	2 characters
Maximum number of characters in an inquiry e.g. 我 正 在 “北 京 郵 電 大 學”。 從 “這 裡”， 我 想 依 次 到 “北 京 航 空 航 天 大 學”、 “中 國 地 質 大 學”、 “北 京 科 技 大 學” 和 “北 京 醫 科 大 學”， 可 以 選 擇 什 麼 交 通 路 線？ <i>I'm at the "Beijing University of Post and Telecommunications". From "here", I want to go to the "Beihang University", "China University of Geosciences", "University of Science and Technology Beijing" and "Beijing Medical University" in sequence. What are the routes available?</i>	54 characters

Table 3.6: Details of the multimodal corpus collected.

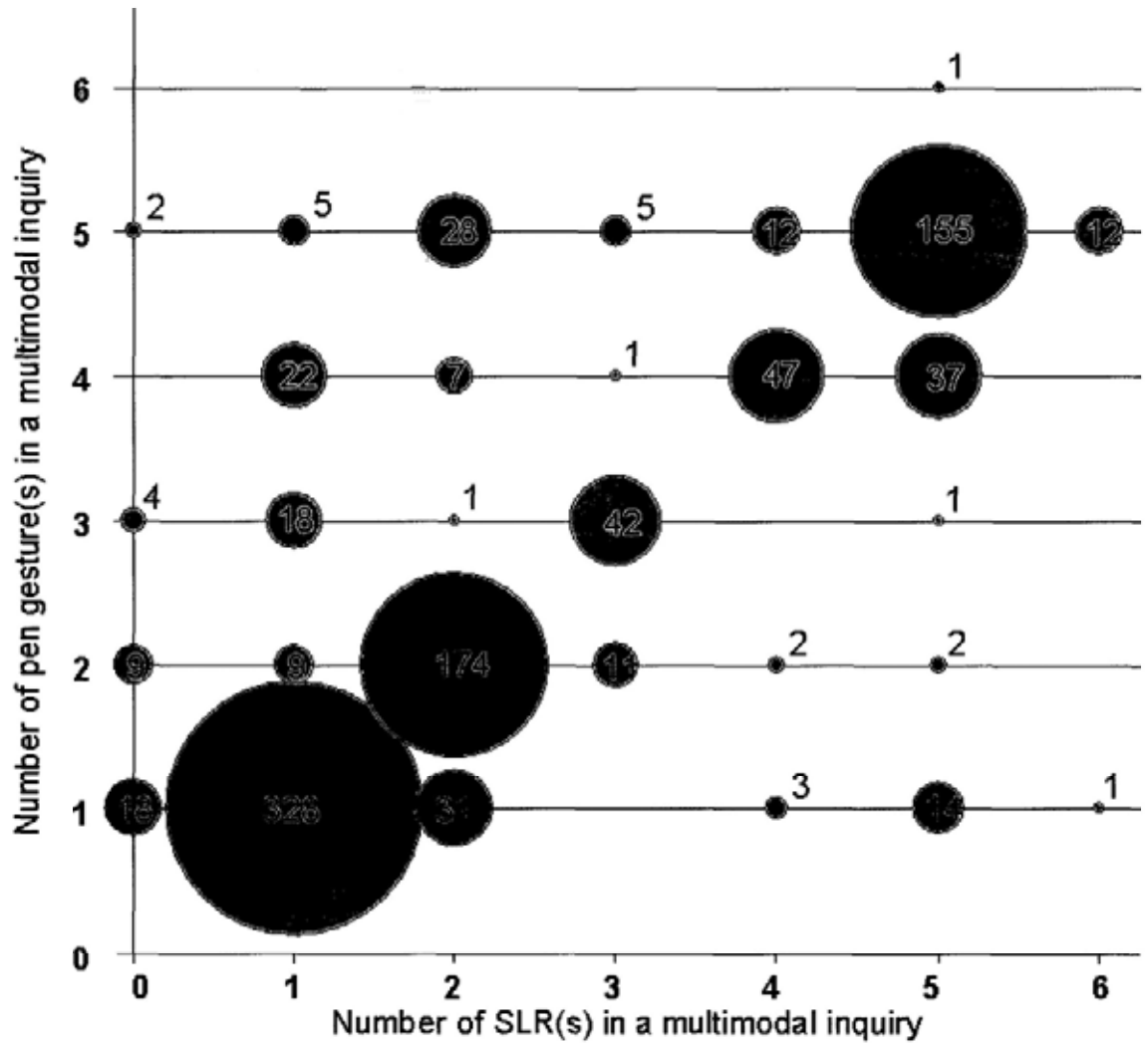


Figure 3.4: The distribution of multimodal inquiries with different numbers of non-spurious SLRs and pen gestures (i.e. manual transcriptions) in the training set. The number inside a bubble is the number of multimodal inquiries with that particular number of SLRs and pen gestures.

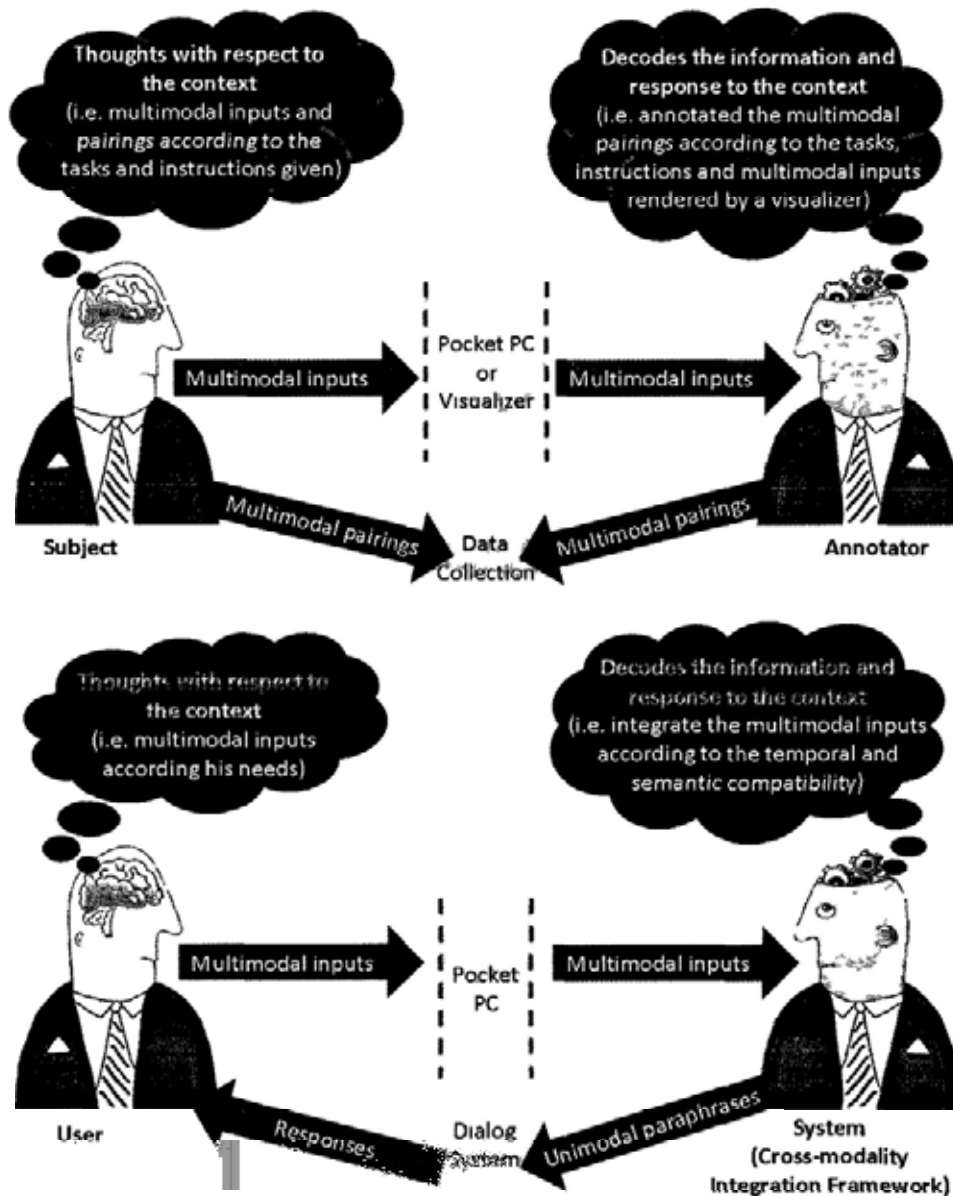


Figure 3.5: An illustration on the nature of the annotator.

We have made the following assumptions during the manual annotation

- The speech and pen inputs are in temporal order. This is true as all of the subjects are cooperative subjects
- Each of the point and stroke is labeled with the location name of its nearest icon in distance
- A circle can be interpreted as multiple locations, i.e. all the icons and labels that “overlap” with the area of the circle
- Each of the SLRs are labeled with the location(s) given in the task (i.e. in the instruction) whenever available. Otherwise, it is labeled with the location(s) with the same location type and subtype
- It is possible to align a single SLR with zero, one or multiple pen gestures and vice versa

Manual annotation follows the steps below

- Ignore disfluencies in the speech modality (e.g. filled pauses and repairs) and spurious gestures in pen modality (e.g. due to jittery hands). In this example with a speech repair, i.e. 我想從這裡到這裡, 不是 是從這裡到這裡 *I want to go from here to here. No. Should be from here to here*, we will only consider the utterance after “no” (i.e. two “這裡” *here* instead of four)
- Record available contextual information (i.e. the current location)
- If the SLR refers to starting location (e.g. 從 *xx from xx*, 由 *xx from xx*, 由 *xx 開始 begin with xx*, etc.), the annotator will first look for contextual information of “current location”. Otherwise, it is labeled as the current location mentioned in the same inquiry

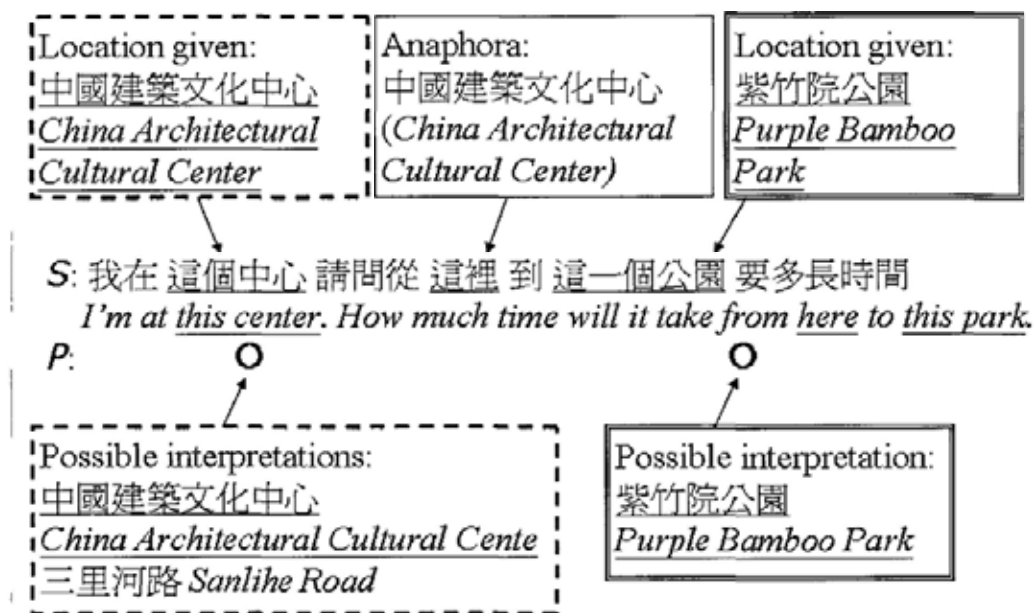
- Compare the location(s) referred by a SLR as given in the task with each of the unpaired pen gesture.
- Pair up the SLR and unpaired pen gesture if the locations referred by them match. Continue to compare the locations referred by both modalities until the number of locations referred by the SLR is satisfied.
- Leave the unpaired pen gesture alone and continue with the next one if none of the locations referred by both modalities match.

The fulfillment of the number of locations referred is necessary for correct alignment since a SLR can correspond with multiple pen gestures (and vice versa). Figure 3.6 is an illustration of the manual annotation process.

Comparison between the annotation obtained with the subject's initial intention (i.e. the pairing indicated by the subject during data collection) shows that there can be multiple possible multimodal pairings of the same inquiry that convey the same meaning. A possible extension on this work can be on the analysis of mismatch between subject's multimodal pairings (i.e. subject's initial intention) and annotator's multimodal pairings.

### 3.6 Chapter Summary

In this chapter, we present the process of design and collection of a multimodal corpus with speech and pen gestures. The corpus is a collection of 1,518 multimodal navigational inquiries around the Beijing area. The speech and pen data of the corpus is manually transcribed. We have also manually annotated the domain-specific named entities and SLRs in the transcribed speech and manually annotated the cross-modality pairings between an SLR from speech and a pen gesture. An SLR may map to zero, one or multiple pen gesture(s) and vice versa. With the multimodal corpus, we can analyze the characteristics of each modality, their relationship and how should they be integrated.



Multimodal Pairings: (1,1) (2,) (3,2)

Figure 3.6: An illustration of the process of manual annotation of cross-modality pairings. The input contains three SLRs which correspond to two locations on the map, together with two circling pen gestures. The locations in the boxes are the possible interpretations of an input event. The boxes with the same border are paired up and the underlined locations are the matched locations between the paired, cross-modal events. The indices of SLRs and pen gestures begin at zero. (1,1) means that the first SLR is aligned with the first pen gestures. In this example, the second SLR “here” does not align with any pen gesture so its pairing is labeled as (2,).



## Chapter 4

# Unimodal and Cross-modal Characterizations

This section describes our findings in an exploratory data analysis of the collected multimodal corpus. Unimodal characterization is referring to characterizing speech and pen gestures and cross-modal characterization is referring to the cross-modality associations between speech and pen gestures. Our aim is to understand how individual modalities encode partial semantics that should later be conjoined to decode the holistic meaning of the user's multimodal input. Results from the analysis are used to devise unimodal interpretation strategies for individual modalities. We have also analyzed the associations between the two modalities, which include the correspondence between modalities and their temporal relationships. According to the characteristics of spoken locative references (SLRs), pen gestures and their temporal relationships, we can design the format of a multimodal term, which is used to represent the cross-modality integration patterns adopted by the user.

Table 4.1 shows the overview statistics of the multimodal corpus collected.

Number of unimodal inquiries (speech only): 76	
Number of inquiries with spoken locative references e.g. 我要看整個“海澱區” <i>I wish to look at the whole “Haidian District”</i>	9 (11.8%)
Number of inquiries without spoken locative references e.g. 我想坐公交車 <i>I wish to take the bus.</i>	67 (88.2%)
Number of multimodal inquiries: 1442	
Number of inquiries with spoken locative references e.g. 從“這個文化中心” <point> 到“這個公園” <point> 要多久? <i>How long does it take to travel from “this center” &lt;point&gt; to “this park” &lt;point&gt; ?</i>	1402 (97.2%)
Number of inquiries without spoken locative references e.g. <stroke> 最快怎麼走? <i>&lt;stroke&gt; What is the fastest route?</i>	40 (2.8%)

Table 4.1: Overview statistics of the Multimodal Corpus. Translations are italicized.

## 4.1 Characterization of Spoken Inputs

The collected data offers over 3,421 (count by token) and 177 (count by type) occurrences of spoken locative references (SLRs) for analysis, from which we can characterize the SLRs in two different ways:

- by the referent of SLRs
- by the numeric feature of SLRs

We can derive the following characterizations by considering the referent of SLRs:

- (1) **Direct references** These involve the use of the full name of a location (e.g. 北京郵電大學 for *Beijing University of Post and Telecommunications*), its abbreviated name (e.g. 北郵 or *BUPT*), or a contextual phrase (e.g. 目前的所在地, *my current location*). Recall that the subject’s “current location” is indicated by a red cross on the map. There are 1,529 occurrences of direct references involving 76 unique tokens/phrases in our corpus.
- (2) **Indirect references** The user may also refer to a location through deixis or anaphora, e.g. 這裡 *here*, 那個中心 *that center*, 這三個商場 *these three shopping centers*, etc. Hence, indirect references may contain numeric features (as indicated with a numeric expression, e.g. 三 *three*, 幾 *few*, 些 *some*, etc.) and/or location type features (e.g. 公園 *park*, 大學 *university*). Both attributes may also be left unspecified in the SLR (e.g. 地方 *place*, 地點 *location*). The location type feature may also be ambiguous (e.g. 站 *station/stop*). There are 1,892 occurrences of indirect references involving 101 unique SLR expressions in our corpus.

In comparison with previous work, the SLRs corresponds to the Givenness Hierarchy with four cognition statuses as mentioned in [41] (see Section 2.2.2),

where the direct references are the uniquely identifiable referents and the indirect references are the activated or familiar referents. Their distributions are shown in Figure 4.1

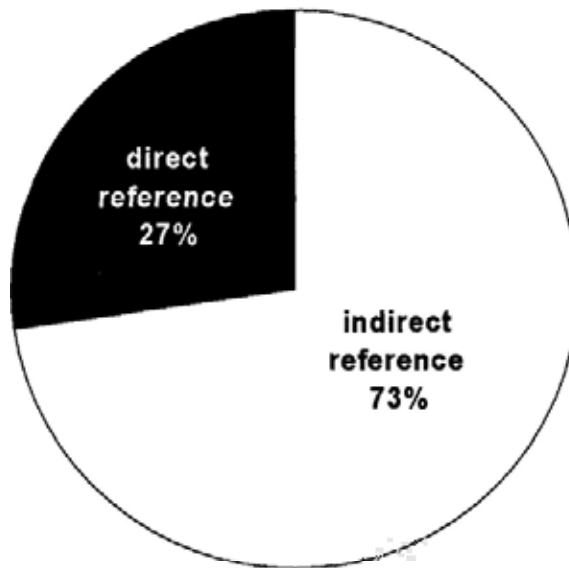


Figure 4.1: Distribution of SLRs according to the types of referent in the training set.

We can derive the following characterizations by considering their *numeric* features:

- (1) **Singular references** A singular reference can be a direct reference with a full name or an abbreviated name. It may also be a singular indirect reference (e.g. 這個公園 *this park*), which may optionally include information about the location type (i.e. a park in the given example).
- (2) **Aggregated references** An aggregated reference is an indirect reference with a specific numeric value (which is greater than one) and an optional location type feature (e.g. 這四個地方 *these four locations*).
- (3) **Plural references** A plural reference is an indirect reference with the numeric feature set to plural (i.e. NUM=plural), as well as an optional

location type feature (e.g. 這些大學 *these universities*).

- (4) **Unspecified references** An unspecified reference is an indirect reference with unspecified numeric and location type features (e.g. 這裡 *here*).

Their distributions are shown in Figure 4.2.

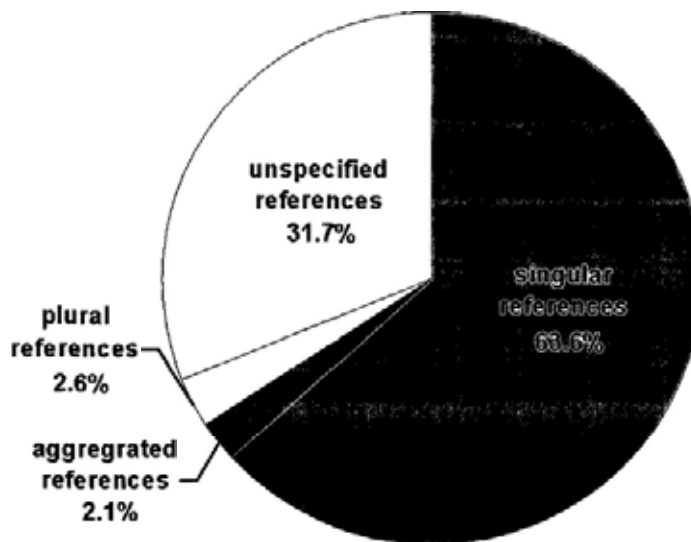


Figure 4.2: Distribution of the types of SLRs according to their numeric features in the training set.

## 4.2 Procedure for Interpreting Spoken Locative References

Based on the above observations, we devise a three-step strategy for interpreting transcribed spoken inputs. These can be applied on manual as well as automatic transcriptions of speech.

**Step 1. Chinese word tokenization** The Chinese language does not have an explicit word delimiter. We perform word tokenization using a greedy algorithm with a home-grown Chinese lexicon with 43,000 entries, cov-

ering nouns, verbs, phrases and SLR expressions. Should speech recognition transcripts be used, the SLR should already be tokenized based on the recognizer’s vocabulary, but may be re-tokenized by the current procedure.

**Step 2. SLR Extraction** We extract the SLR expressions by referring to our lexicon, which includes 177 unique SLR expressions. The extraction algorithm also parse the numeric expression and location type expression from the parsed SLR expression. The parsed numeric expression and location type expression are used to fill in the numeric feature attribute and location type feature attribute of the SLR respectively.

**Step 3. Hypotheses generation** This step generates a hypothesized *list of locations* corresponding to a given SLR. A single location is typically generated for direct references, based on the name of the location or the current location from context. The list of hypothesized locations generated for an indirect reference typically includes all icons present on the map. This list may be narrowed down according to a matching location type, if the feature is specified. Furthermore, if the numeric feature is specified, it is stored along with the generated hypothesis list. Rank ordering of the hypothesized locations is not considered for SLRs.

### 4.3 Characterization of Pen Inputs

The training set of our corpus contains 2,564 pen gestures in total. Of these, 1805 (70.4%) are pointing gestures (POINT), 470 (18.3%) are circling gestures (CIRCLE) and 289 (11.3%) are strokes (STROKE). Analysis of the corpus also sheds light on the usages of the different pen gestures as illustrated in Table 4.2.

- (i) POINT This is mostly used to indicate a single location. This occurs 99.8% (1801/1805) of the time in our corpus and the remaining occur-

rences are map rendering commands.

- (ii) **CIRCLE** This includes two possible cases - small circles indicate a single location (70% of corpus statistics, 329/470) and large circles indicate multiple locations (30% of corpus statistics, 141/470).
- (iii) **STROKE** These include three possible cases - a stroke referring to a street or bridge (45.3% of corpus statistics, 131/289), the start and end points of a path (32.4%, 94/289) and multiple strokes constituting a route (22.3%, 64/289).

Analysis of the training set shows that 95% of the multimodal terms contain a single pen gesture, i.e. **POINT**, **CIRCLE** or **STROKE**. The remaining multimodal inputs (i.e. 5%) contain multiple pen gestures, to which we refer as **MULTI-POINT**, **MULTI-CIRCLE** and **MULTI-STROKE**. Table 4.3 shows examples of pen gestures and their semantics.

#### 4.4 Interpreting Pen Inputs

Pen inputs are interpreted based on the gesture type and its coordinates, which are compared with the positional coordinates of the icons on the map. Interpretation of each gesture type generates a ranked hypothesis list of locations, according to the following protocol:

- (i) **POINT**: icons lying within 42 pixels from the point are considered possible semantic interpretations of the gesture. These are ranked according to distances away from the point. Shorter distances are given higher ranks.
- (ii) **CIRCLE**: the circle's area is defined by the pair of coordinates corresponding to the pen-down and pen-up gestures. Icons with overlapping areas are considered possible semantic interpretations and are ranked according to their distances away from the estimated center of the circle.



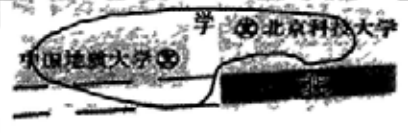
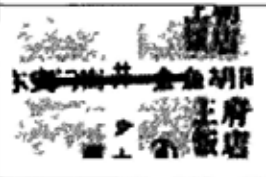

Gesture	Semantics	Illustration(s)
POINT	Indicates a single location, NUM=1, e.g. a university	
CIRCLE	A small circle indicates a single location, NUM=1, e.g. a park	
	A large circle indicates multiple locations, NUM=plural, e.g. 2 universities	
STROKE	A single stroke indicates a single location, NUM=1, e.g. a street	
	A single stroke indicates the start and end points of a path, NUM=1	

Table 4 2. Illustrations of the usages of different pen gesture types.






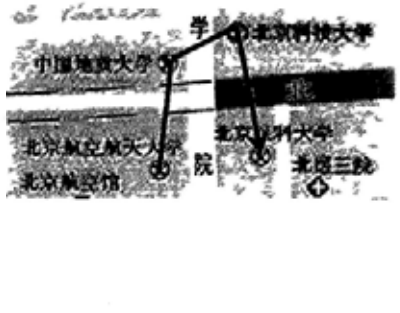
Gesture	Semantics	Illustration(s)
MULTI-POINT	Indicates multiple locations, NUM=1, e.g. four points correspond to one SLR	
MULTI-CIRCLE	Indicates multiple locations, NUM=1, e.g. four circles correspond to one SLR	
MULTI-STROKE	Indicates multiple locations, NUM=1, e.g. three strokes correspond to one SLR	
	Multiple strokes or a long stroke with one or more turning points to indicate a route, NUM=1, e.g. a long stroke passing through four universities	

Table 4.3: Illustrations of the usages of multiple pen gestures.

Again, shorter distances are given higher ranks.

- (iii) **STROKE**: a hypothesis list is generated for each endpoint of a stroke, where hypotheses are ranked by their distances from the endpoint. If we compare the hypothesis list of two adjacent endpoints (from one stroke or two sequential strokes) and find significant similarity (i.e. either the top three entries are identical, or the two lists have over 75% overlap), the two hypothesis lists will be merged into one according to their common entries. Using this method, we can distinguish between interpreting a single stroke as one location, from the other alternative of a connecting stroke between two locations. In the case of multiple sequential strokes, such as the three strokes in Table 4, this method enables us to interpret them as a route connecting four locations.

Table 4.4 illustrates the process of interpreting speech and pen gestures interpretation procedure.

## 4.5 Temporal Relationships

As mentioned in Section 3.5, we annotate the correspondence between SLRs and pen gestures based on temporal ordering and semantic compatibility (i.e. type and the number of location(s) referred). Since there can be one-to-many mapping between the SLR and its associated pen gestures, the pen gestures are considered together as a group (i.e. **MULTI-POINT**, **MULTI-CIRCLE** or **MULTI-STROKE**). The reverse is also true when mapping a pen gesture to multiple SLRs. Analysis of the training data shows that in a multimodal input, SLR and pen gesture that (jointly) refer to the same intended location may not always overlap in time.


As observed in our training set, temporal integration patterns [17] between corresponding SLRs and pen gestures include two main types: simultaneous

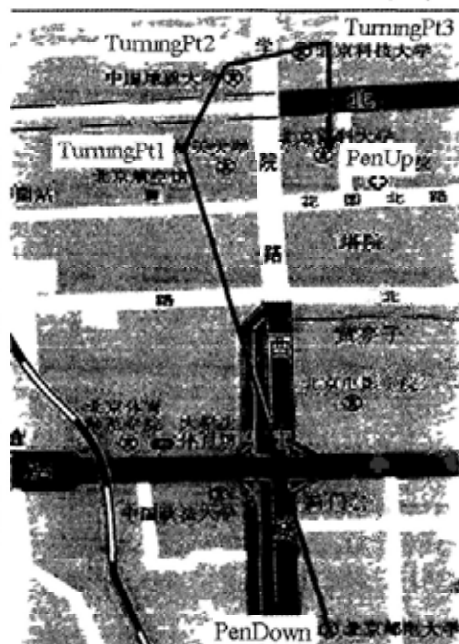
**Multimodal input**

S. 我現在在“北郵”我要到“這四個大學”一共需要多少時間

*I am now at “BUPT” and I need to get to “these four universities”.*

*How much time will it take?*

P:  (a long stroke with three turning points)

**Hypothesis lists of speech input**

SLR1: ABBREVIATION=北郵 BUPT

北京郵電大學 *Beijing University of Posts and Telecommunications*

SLR2: DEICTIC=這四個大學 *these four universities*

NUM=4

LOC\_TYPE=SCHOOLS AND PUBLIC LIBRARIES

subtype=*university*

中國地質大學

北京師範大學

北京郵電大學 .....(all universities on the map shown)

*continue...*

Hypothesis lists of pen input ( <i>locations ranked by distance in pixels</i> )		
PenDown: TYPE=STROKE		
北京郵電大學	-1	<i>Beijing University of Posts and Telecommunications</i>
西土城路	5.4	<i>Xitucheng Road.....</i>
TurningPt1: TYPE=STROKE		
北京航空航天大學	-1	<i>Beihang University</i>
北京航空館	5.0	<i>Beijing Aviation Museum.....</i>
TurningPt2: TYPE=STROKE		
中國地質大學	1.9	<i>China University of Geosciences</i>
學院路	11.0	<i>Xueyuan Road.....</i>
TurningPt3: TYPE=STROKE		
北京科技大學	0.6	<i>University of Science and Technology Beijing</i>
學院路	11.4	<i>Xueyuan Road.....</i>
PenUp: TYPE=STROKE		
北京醫科大學	-1	<i>Beijing Medical University</i>
北醫三院	7.02	<i>Peking University Third Hospital.....</i>

Table 4.4: An illustration of the procedure for hypothesis lists generation in the speech and pen modalities respectively. Translations are italicized. Distance labeled with “-1” means the pen gesture is triggered inside the area of icon/label of that location.

(SIM) and sequential patterns (SEQ). Simultaneous SLRs and pen gestures have temporal overlap between an SLR and its corresponding pen gesture(s) (no matter when is the start/end time). Sequential associations do not have temporal overlap between the duration of SLR and its corresponding pen gesture. A 3-tuple that consists of corresponding SLR(s) and pen gesture(s), together with their temporal relationship, i.e. <SLR | pen\_gesture\_type | temporal\_relationship> is referred as a multimodal term. Among the 2261 multimodal terms found in the training set, 74% are simultaneous and 26% are sequential. For example, consider the multimodal expression:

S: 從“我所在的地方”到“這裡”可以怎麼走？

*How can I go from “my current location” to “here”?*

P:                   •                   ••••

Since the four points are considered as a group as MULTI-POINT and temporally overlapped with the SLR 這裡 *here*, the temporal relationship between them is simultaneous. Therefore, the multimodal terms of this multimodal expression include <我所在的地方 | POINT | SIM> and <這裡 | MULTI-POINT | SIM>.

Further classification of simultaneous input patterns shows that there are nine logically possible overlap patterns [17]. Statistics of the nine overlap patterns are shown in Table 4.5. Input patterns with speech showed temporal precedence (the third column of Table 4.5) accounts for the majority (i.e. 87.68% of the total).

Sequential patterns can be further classified into two: speech precedes pen (72.5%) and vice versa with pen precedes speech (27.5%). Statistics are shown in Table 4.6. In this work, the maximum lag time between speech and pen is around seven seconds and the distribution of the lag time of the two sequential patterns (i.e. speech precedes pen and vice versa) are shown in Figures 4.3 and 4.4 respectively. It shows that around 80% of the sequential inputs have lag ranging between zero and one second.



















Neither Precedes (2.09%)	Pen Precedes (10.23%)	Speech Precedes (87.68%)
<b>S:</b>  <b>P:</b>  (0%)	<b>S:</b>  <b>P:</b>  (0.63%)	<b>S:</b>  <b>P:</b>  (1.25%)
<b>S:</b>  <b>P:</b>  (0%)	<b>S:</b>  <b>P:</b>  (1.25%)	<b>S:</b>  <b>P:</b>  (19.21%)
<b>S:</b>  <b>P:</b>  (2.09%)	<b>S:</b>  <b>P:</b>  (8.35%)	<b>S:</b>  <b>P:</b>  (67.22%)

Table 4.5: Nine logically possible temporal overlap patterns between speech and pen gestures for simultaneous inputs.





Speech Precedes (72.5%)	Pen Precedes (27.5%)
<b>S:</b>  <b>P:</b>  lag time	<b>S:</b>  <b>P:</b>  lag time

Table 4.6: Two temporal patterns between speech and pen gestures for sequential inputs.

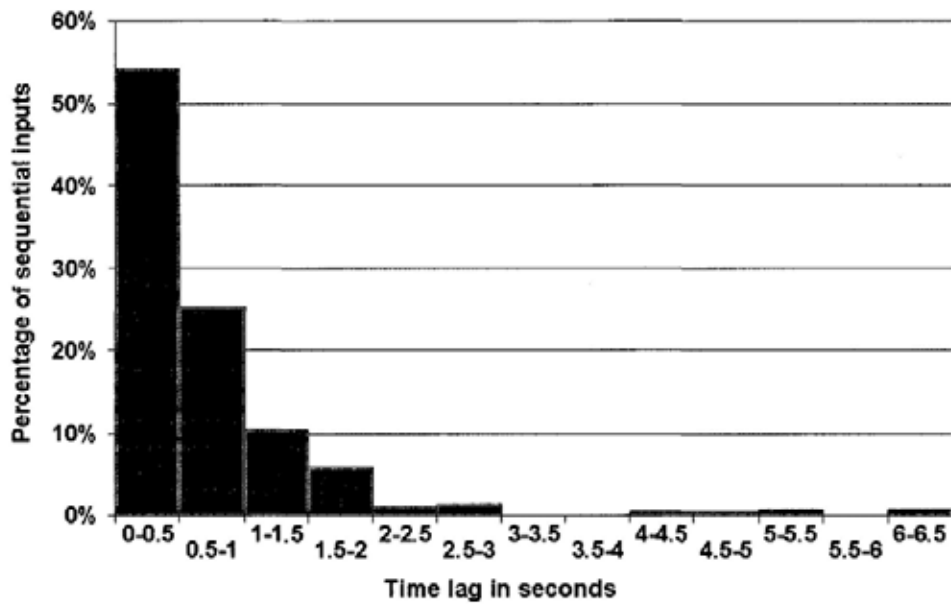


Figure 4.3: Distribution of lag times between end of speech and onset of pen (i.e. speech precedes) in sequential inputs.

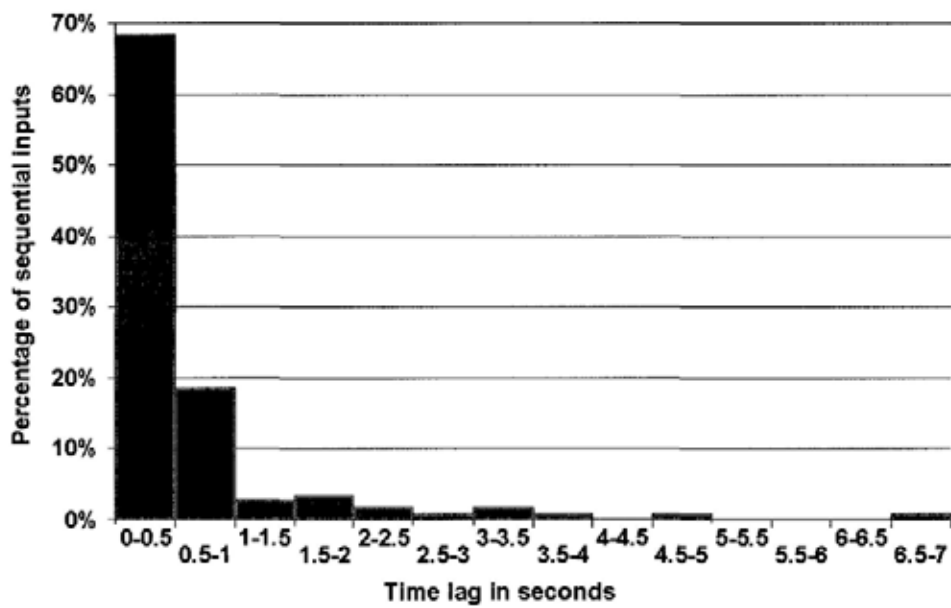


Figure 4.4: Distribution of lag times between end of pen and onset of speech (i.e. pen precedes) in sequential inputs.

Further analysis on the subjects' multimodal input integration patterns shows that most of the subjects have adopted either a dominantly simultaneous or sequential temporal patterns between speech and pen gestures [59] [60]. Table 4.7 shows the statistics of subjects' temporal patterns. Statistics show that subjects have a relatively high consistency (82.6%) in their integration patterns. These findings are important for further development of the proposed framework so that it will be able to adopt to subject's dominant integration pattern.

## 4.6 Cross-Modal Integration Patterns

As mentioned in Section 4.1, SLRs may be singular, aggregated, plural and unspecified references. Recall that an SLR may correspond to one or more pen gestures. We analyze the statistics in the training set as shown in Table 4.8. From the statistics, we observe that users predominantly prefer to use a single reference in the SLR (62.5%, as shown in the first row of Table 4.8). Furthermore, a single SLR generally corresponds to a single pen gesture, as none were found mapping to multiple pen gestures.

As regards aggregated references (e.g. 這四個大學 *these four universities*), 78.6% were found to correspond with multiple pen gestures to indicate multiple locations. The other 16% are used with a circle (i.e. a single pen gesture) that encompasses multiple locations. An example is the multimodal term <這四個大學 | CIRCLE | SIM > or <*these four universities* | CIRCLE | SIM>.

For plural references, as shown in the third row of Table 4.8, 72% are used with multiple pen gestures to indicate multiple locations. The remaining 28% are used with a single pen gesture, with the majority (19/21) being circles and the remaining two are points.

SLRs with an unspecified numeric features should correspond to both single and multiple pen gestures. Within the training set, however, an unspecified



Subject	SIM	SEQ
Subjects with dominant simultaneous integration pattern		
1	96.69%	3.31%
2	86.96%	13.04%
3	74.49%	25.51%
4	99.05%	0.95%
5	73.61%	26.39%
6	89.36%	10.64%
7	69.62%	30.38%
9	76.27%	23.73%
10	74.70%	25.30%
11	71.91%	28.09%
12	74.19%	25.81%
14	97.14%	2.86%
15	68.48%	31.52%
16	70.64%	29.36%
18	80.73%	19.27%
19	84.62%	15.38%
20	70.31%	29.69%
21	91.87%	8.13%
22	68.38%	31.62%
23	92.70%	7.30%
Subjects with dominant sequential integration pattern		
13	7.14%	92.86%
17	1.00%	99.00%
Subject without dominant integration pattern		
8	50.00%	50.00%

Table 4.7: Percentage of simultaneous and sequential temporal patterns for all 23 subjects. Average consistency of user's dominant integration pattern is 82.6%.

reference predominantly (94%) occurs in association with a single pen gesture.

The above refers to SLRs that are deictic or anaphoric expressions. Deictic expressions need to be interpreted jointly with the associated pen gestures. Anaphoric references are interpreted based on contextual information and do not correspond to any pen gestures. The first row in Table 4.9 presents examples of these two types of expressions. Additionally, there are also elliptic expressions, where the SLR is completely omitted but the pen gesture is present. For such cases, the cross-modal temporal relationship is irrelevant (and indicated by “ $\emptyset$ ”). Table 4.9 shows some examples.

The number of multimodal terms is much fewer than the exhaustive combinations between SLRs and pen gestures. Some of the terms are not found in the corpus, while others may be implausible combinations, such as:

- A singular reference with multiple pen gestures (e.g. <這個大學 | MULTI-POINT | SIM> <*this university* | MULTI-POINT | SIM>) - a singular SLR refers to one location and corresponds to one pen gesture. Multiple pen gestures should correspond to an aggregated or plural reference. Therefore, this combination involves incompatibility in the numeric feature. This constraint can be used to mutual disambiguation. This is because if is an impossible combo. So if it occurs, probably it means that there is a recognition error. We will make use of this in our work on cross-modality integration (i.e. semantic compatibility on numeric feature which will be describe in Chapter 5).
- An aggregated reference with a single point or a single stroke (e.g. <這三個地方 | POINT | SIM> <*these three places* | POINT | SIM>) - an aggregated SLR refers to multiple locations and should correspond to multiple pen gestures or a circle. Again, this combination involves incompatibility in the numeric feature.
- An unspecified reference with multiple circles or strokes (e.g. <這裡 |

Speech (as parsed SLR from transcribed speech)	Pen (as transcribed gesture)	Temporal Relationship (SIM / SEQ)	Count
<b>Singular</b> (1550/2480, 62.5%)	<b>Single</b> (1417/1550, 91.4%)	SIM (1024/1417, 72.3%)	1024
		SEQ (393/1417, 27.7%)	393
	<b>Multiple</b> (0/1550, 0%)	SIM	0
		SEQ	0
	$\emptyset$ (133/1550, 8.6%)	$\emptyset$	133
<b>Aggregated</b> (56/2480, 2.3%)	<b>Single</b> (9/56, 16%)	SIM (7/9, 77.8%)	7
		SEQ (2/9, 22.2%)	2
	<b>Multiple</b> (44/56, 78.6%)	SIM (25/44, 56.8%)	25
		SEQ (19/44, 43.2%)	19
	$\emptyset$ (3/56, 5.4%)	$\emptyset$	3
<b>Plural</b> (75/2480, 3%)	<b>Single</b> (21/75, 28%)	SIM (12/21, 57.1%)	12
		SEQ (9/21, 42.9%)	9
	<b>Multiple</b> (54/75, 72%)	SIM (35/54, 64.8%)	35
		SEQ (19/54, 35.2%)	19
	$\emptyset$ (0/75, 0%)	$\emptyset$	0
<b>Unspecified</b> (761/2480, 30.7%)	<b>Single</b> (715/761, 94%)	SIM (569/715, 79.6%)	569
		SEQ (146/715, 20.4%)	146
	<b>Multiple</b> (1/761, 0.1%)	SIM (1/1, 100%)	1
		SEQ (0/1, 0%)	0
	$\emptyset$ (45/761, 5.9%)	$\emptyset$	45
$\emptyset$ (38/2480, 1.5%)	<b>Single</b> (34/38, 89.5%)	$\emptyset$	34
	<b>Multiple</b> (4/38, 10.5%)	$\emptyset$	4

Table 4.8: Statistics of cross-modal integration patterns in the training set. There are altogether 2480 multimodal terms (count by token) in total. Among them, 2261 contain both SLR and pen gesture, 181 contain only SLRs and 38 of them contain only pen gestures.

<p><b>User input with deictic and anaphoric references (the second “here” is an anaphora to the first “here”):</b></p> <p>S: 我在 “這裡” 從 “這裡” 到 “這裡” 要 多久</p> <p>P:           •                           ○</p> <p><i>I’m now “here.” How much time will it take to go from “here” to “here”?</i></p>
<p><b>Annotated user input with multimodal terms:</b></p> <p>我在 &lt;這裡   POINT   SIM&gt; 從 &lt;這裡   ◊   ◊&gt; 到 &lt;這裡   CIRCLE   SEQ&gt; 要 多久</p> <p><i>I’m now at &lt;here   POINT   SIM&gt;. How much time will it take from &lt;here   ◊   ◊&gt; to &lt;here   CIRCLE   SEQ&gt;?</i></p>
<p><b>User input with elliptic locative references (the SLR is omitted in speech):</b></p> <p>S:       開放時間 Opening hours?</p> <p>P •••</p>
<p><b>Annotated user input with a multimodal term:</b></p> <p>&lt;◊   MULTI-POINT   ◊&gt; 開放時間</p> <p>&lt;◊   MULTI-POINT   ◊&gt; <i>Opening hours?</i></p>

Table 4.9: Examples on 3-tuple multimodal term annotation with speech (S) and pen gesture (P). Translations are italicized.

MULTI-STROKE | SIM> <here | MULTI-STROKE | SIM>) - empirically, we have found that about 94% of the unspecified references are used to indicate a single location (as shown in Table 4.8). A possible reason may be that unspecified SLRs have short durations, during which the subjects may find it difficult to gesture multiple circles or strokes simultaneously.

## 4.7 Chapter Summary

In this chapter, we present the characteristics of spoken and pen inputs. Spoken inputs can be categorized as direct or indirect references according to their referents or as singular, plural, aggregated or unspecified references according to their numeric features. We have devised a processing sequence for extracting SLRs from the manually transcribed spoken input and interpreting each SLR by generating a hypothesis list of possible semantics (i.e. locations). Pen inputs can be illustrated as point, circle or stroke (and also multiple occurrences of each pen gesture). We have also devised a processing sequence for interpreting pen gestures and generating a hypothesis list for every gesture. We have analyzed the temporal patterns between the two modalities and found that there are two main types of temporal integration patterns, including simultaneous and sequential patterns. The majority (74%) of the temporal relationships found are simultaneous pattern. For the sequential pattern, the maximum time lag speech speech and pen gestures is seven seconds. Statistics also show that over 95% of the subjects (22/23) have their dominant integration pattern and the average consistency of user's dominant integration pattern is 92.6%. According to the characteristics of SLR, pen gesture and their temporal relationship, we have designed a representation for multimodal term using a 3-tuple, which consists of an SLR, pen gesture and their temporal relationship. Such multimodal terms is used to represent the cross-modality integration patterns adopted by the user.

## Chapter 5

# Cross-Modality Semantic Integration

As described in Chapter 4, each of the two (speech and pen) modalities abstracts the user's intended message into a sequence of input events, i.e., in terms of spoken locative references (SLRs) or pen gestures. Each event carries semantic meaning but may contain ambiguity. The interpretation procedures for speech and pen inputs presented in the previous chapter (Sections 4.2 and 4.4) derive partial semantics for each event, represented as a hypothesized list of locations. This chapter presents a cross-modality integration procedure that attempts to integrate the partial interpretations across modalities in order to generate a unimodal paraphrase that is semantically equivalent to the original multimodal user input. We perform the cross-modality integration by Viterbi alignment which enforces the constraints of temporal order and semantic compatibility constraints between speech and pen gestures. We apply the cross-modality integration procedure on manual transcription (i.e. perfect recognition outputs) and top-scoring automatic recognition outputs so as to obtain the upper and lower bound of the integration performance. Besides, in order to gain an empirical understanding of the inter-relations between the

speech and pen modalities, this chapter also presents a comparative analysis of multimodal user inputs with their generated, semantically equivalent unimodal paraphrase based on class trigram perplexities. Analysis shows two categories of data (i.e.  $PP_{MM} < PP_{UM}$  and  $PP_{MM} = PP_{UM}$ ) will also be described in this chapter.

## 5.1 Cross-Modality Integration on Perfect Transcriptions

Statistics in Section 3.4 show that around 74.5% (746/1002) of the multimodal inquiries in the training set have an equal number of SLR and pen gestures. However, in these cases, there may not be a one-to-one correspondence between the SLRs and pen gestures. For example

*S* 從“所在地”到“這兩個地方”要多久

*P* ● ●

*How long will it take to travel from “my current location” to “these two locations”?*

There are two SLRs and two pointing gestures in the inquiry. However, the first SLR is an anaphora referring to the user’s current location, and the two pointing gestures both correspond to the second SLR. If we consider only the inquiries with SLR(s) in the training set, there are 968 (out of 1002) multimodal inquiries contain both SLR(s) and pen gesture(s). An overly bold assumption of one-to-one correspondence between SLRs and pen gestures can correctly interpret only around 67.3% (651/968) of the perfectly transcribed multimodal inquiries in the training set.

Statistics in Section 4.5 shows that 74% (1673/2261) of the multimodal

terms in the training set have simultaneous timing relationships between corresponding SLRs and pen gestures. The remaining 26% (588/2261) have sequential timing with a maximum time lag of seven seconds. If we simply look for a pen input that occurred closest in time to the SLR [61], this can only correctly interpret 75.1% (727/968) of the perfectly transcribed multimodal inquiries (which contain SLR(s)) in the training set.

Therefore, we perform cross-modality integration by Viterbi alignment [62] with a scoring function that enforces the temporal ordering between the sequence of SLRs and the sequence of pen gestures. The scoring function also enforces the semantic compatibility in terms of numeric (NUM) and location type (LOC\_TYPE) features (see Figure 5.1).

### 5.1.1 Enforcing Temporal Order

Analysis of our training data shows that in a multimodal input expression, the spoken locative reference (SLR) and pen gesture that correspond to the same intended location may not always overlap in time. In fact, about one-fourth (see Section 4.5) of cases in the training set show the pen gesture occurring either before or after its corresponding spoken reference (i.e. sequential inputs). Hence in the current work, we only attempt to maintain the temporal order of locative references between the speech and pen inputs. A Viterbi alignment  $a = a_1 a_2 a_3 \dots a_m$  can easily accommodate for this as we align the sequence of  $R$  hypothesis lists in temporal order of the SLRs  $S = S_1 S_2 \dots S_R$  with the sequence of  $Q$  hypothesis lists in temporal order of the pen gestures  $P = P_1 P_2 \dots P_Q$ . Note that it is possible for a single SLR to align with multiple pen gestures (e.g. “*these three universities*” is a single SLR that corresponds to three pointing inputs), as well as vice versa (e.g. “*Xueyuan Road and North Huyuan Road*” corresponds to one circling gesture). The Viterbi alignment algorithm can support this by advancing the position in one hy-



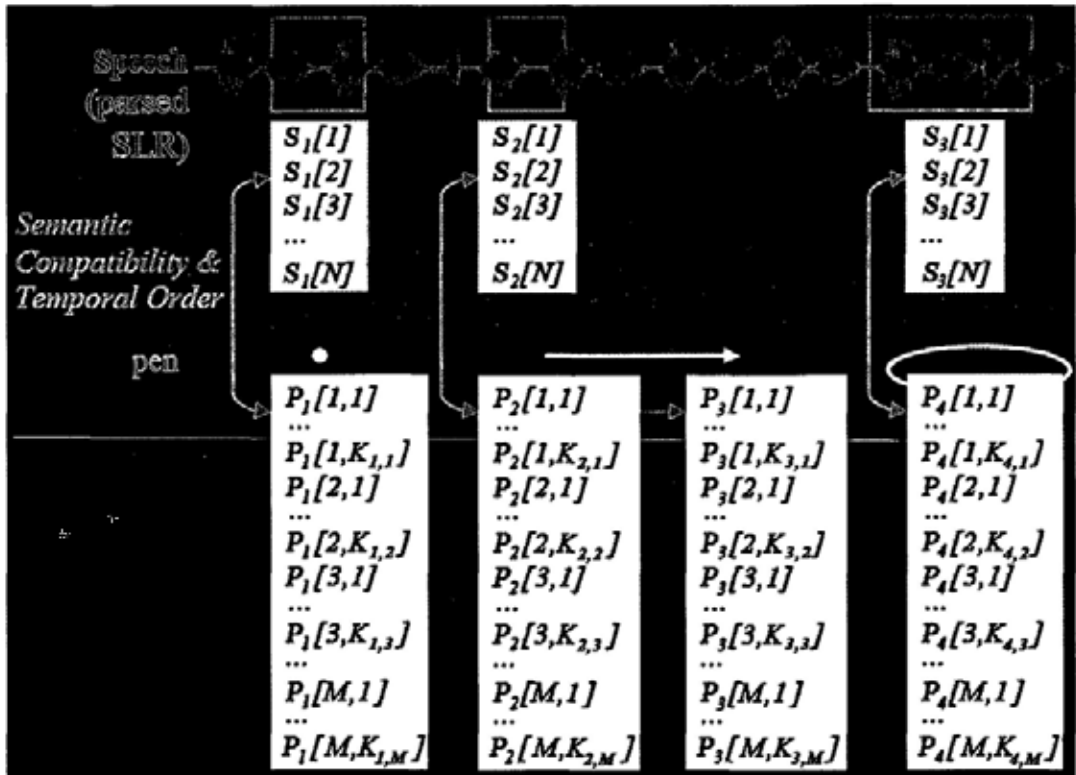


Figure 5.1: The cross-modality integration procedure. Each input event (a spoken locative reference or a pen gesture such as POINT/CIRCLE/STROKE) in each modality produces a list of hypothesized locations. These are aligned across modalities by the Viterbi algorithm while incorporating semantic compatibility and temporal order.  $S_r[N]$  is the  $N$ -best recognition hypothesis of the  $r^{th}$  SLR and  $P_q[M, K_{q,M}]$  is the  $M$ -best recognition hypothesis of the  $q^{th}$  pen gesture instance with  $K_{q,M}$  hypothesized locations.

pothesis sequence (either  $S_r$  to  $S_{r+1}$  or  $P_q$  to  $P_{q+1}$  as indicated in Figure 5.1) while maintaining the position in the other.

### 5.1.2 Enforcing Semantic Compatibility

Cross-modality integration also seeks to enforce semantic compatibility. If the  $r^{\text{th}}$  SLR is a direct reference expression, the hypothesis list  $S_r$  should contain only one element and the integration procedure seeks to match the specified location with hypotheses for the aligned pen gesture in  $P_q$ . The matching cost is defined such that if no match is found, a cost of one is incurred. If the SLR is an indirect reference expression, the hypothesis list  $S_r$  should contain multiple elements and the location type (LOC\_TYPE) or numeric (NUM) features may be specified. The integration procedure checks for compatible LOC\_TYPE among the hypotheses for the aligned pen gesture in  $P_q$ . A *matching cost*  $C_M(S_r, P_q)$  of one is incurred if there is mismatch in LOC\_TYPE between  $S_r$  and  $P_q$  (see Equation 5.3 in Table 5.1). Enforcing compatibility in NUM is a little more elaborate, especially when the value of NUM specifies multiple locations that need to be matched with the hypothesis sequences from recognized pen gestures. Hence we use a *transition cost*  $C_T(S_r, P_q | S_{r-1}, P_{q-j})$  which is set to the deficit/excess in the NUM value during the transition from  $(S_r, P_{q-1})$ ,  $(S_{r-1}, P_q)$  or  $(S_{r-1}, P_{q-1})$  to  $(S_r, P_q)$  as shown in Equation 5.4 (see Table 5.1). This is used to indicate that there are too few or too many pen gestures aligned with one SLR or vice versa. The matching cost of location type and transition of numeric feature are determined with the training set. As mentioned, an SLR may align with one or more pen gestures, corresponding to one or more  $P_q$  and each may contain a different number of hypotheses. Should we encounter a tie in the conditional cumulative costs  $C_C(S_r, P_q | S_{r-1}, P_{q-j})$  at  $(S_r, P_q)$  from different positions  $(S_{r-1}, P_{q-j})$  during the course of alignment, we pick the back pointer  $B(S_r, P_q)$  in the following order of precedence:

1. Return one step in  $S_r$  while maintaining the position in  $P_q$  (i.e.  $i = 1$  and  $j = 0$  leading to  $S_{r-1}$  and  $P_q$ ). This order aims to handle the occurrence of anaphoric reference to the user's existing location, i.e. the anaphora does not need to pair up with a pen gesture.
2. If the above path is not available, return one step in both  $P_q$  and  $S_r$  (i.e.  $i = 1$  and  $j = 1$  leading to  $S_{r-1}$  and  $P_{q-1}$ ). This order aims to handle the one-to-one correspondence between speech and pen gestures - around 67.2% of the SLR has one-to-one correspondence with pen gesture and can be correctly interpreted.
3. If the above path is not available, return one step in  $P_q$  while maintaining the position in  $S_r$  (i.e.  $i = 0$  and  $j = 1$  leading to  $S_r$  and  $P_{q-1}$ ).

Details of the Viterbi algorithm are provided in Table 5.1. An illustrative example is shown in Figure 5.2.

### 5.1.3 Identifying Intended Locations

This alignment procedure generates the “best” path in attempting to find an alignment between an SLR with a pen gesture in the multimodal input. The cross-modality integration procedure extracts the common location(s) found in each pair of hypothesis lists ( $S_r$  and  $P_q$ ) derived from the aligned SLR and pen gesture. The number of locations extracted follows the value of the NUM feature and the ranking of locations follows those from the hypothesis list  $P_q$  from the pen modality (as described in Section 4.4). The top ranking location(s) is identified as the user's intended location(s). By substituting the identified locations in place of the SLRs in the speech input, we can generate a *unimodal, verbalized paraphrase* that is semantically equivalent to the original multimodal expression. For example,

**Notations:**

- $S_r$  is the list of hypothesis of the  $r^{th}$  SLR.
- $P_q$  is the list of hypothesis of the  $q^{th}$  pen gesture instance.
- $C_M(S_r, P_q)$  is the *matching* cost between  $S_r$  and  $P_q$ .
- $C_T(S_r, P_q | S_{r-1}, P_{q-j})$  is the *transition* cost from  $(S_{r-1}, P_{q-j})$  to the current position  $(S_r, P_q)$ . It indicates the deficit or exceed in the NUM value for  $i, j = \{0, 1\}$ .
- $C_A(S_r, P_q)$  is the *cumulative* cost (the best partial alignment) up to the position of  $(S_r, P_q)$  from  $(S_1, P_1)$ .
- $C_C(S_r, P_q | S_{r-1}, P_{q-j})$  is the *conditional cumulative* cost at  $(S_r, P_q)$  from the position  $(S_{r-1}, P_{q-j})$  for  $i, j = \{0, 1\}$ , such that  $C_C(S_r, P_q | S_{r-1}, P_{q-j}) = C_M(S_r, P_q) + C_A(S_{r-1}, P_{q-j}) + C_T(S_r, P_q | S_{r-1}, P_{q-j})$ .
- $B(S_r, P_q)$  is the back pointer of the position  $(S_r, P_q)$  determined by the local minimization of  $C_A(S_r, P_q)$ .
- $\Psi(r, q)$  is the backtracking path obtained from the back pointer  $B(S_r, P_q)$ .
- $C_A(S_R, P_Q)$  is the cumulative cost at the final position  $(S_R, P_Q)$ .
- $R$  is the total number of SLRs in the inquiry.
- $Q$  is the total number of pen gesture instances in the inquiry.

**Initialization:**

$$C_A(S_1, P_1) = C_M(S_1, P_1) \quad (5.1)$$

$$B(S_1, P_1) = nl \quad (5.2)$$

**Recursion** ( $\forall(r, q) = \{(1, 1), \dots, (R-1, Q), (R, Q-1), (R-1, Q-1)\}$ ):

$$C_M(S_r, P_q) = \begin{cases} 0 & S_r \cap P_q \neq \emptyset \\ 1 & S_r \cap P_q = \emptyset \end{cases} \quad (5.3)$$

$$C_T(S_r, P_q | S_{r-i}, P_{q-j}) = \text{absolute value of deficit or excess in the NUM value} \\ \text{for } \{i, j\} = \{(0, 1), (1, 0), (1, 1)\} \quad (5.4)$$

$$C_A(S_r, P_q) = \begin{cases} C_C(S_r, P_q | S_r, P_{q-1}) & \text{if } r = 1 \\ C_C(S_r, P_q | S_{r-1}, P_q) & \text{if } q = 1 \\ \min\{C_C(S_r, P_q | S_{r-1}, P_{q-1}), C_C(S_r, P_q | S_{r-1}, P_q), \\ C_C(S_r, P_q | S_r, P_{q-1})\} & \text{otherwise} \end{cases} \quad (5.5)$$

$$B(S_r, P_q) = \arg \min_X \{C_A(S_r, P_q)\} \quad (5.6) \\ \text{for } X = \{(r-1, q), (r, q-1), (r-1, q-1)\}$$

**Termination:**


$$C_A = (S_R, P_Q) \\ = \min\{C_C(S_R, P_Q | S_{R-1}, P_{Q-1}), C_C(S_R, P_Q | S_{R-1}, P_Q), \\ C_C(S_R, P_Q | S_R, P_{Q-1})\} \quad (5.7)$$

$$B(S_R, P_Q) = \arg \min_X \{C_A(S_R, P_Q)\} \quad (5.8) \\ \text{for } X = \{(R-1, Q), (R, Q-1), (R-1, Q-1)\}$$

**Path Backtracking:**

while  $i > 0$ , do  $\{\Psi(r[j], q[j]) := B(S_{r[i]}, P_{q[i]}), i := j\}$   
for  $(r[j], q[j]) = \{(R, Q), \dots, (1, 1)\}$


Table 5.1: Details of the Viterbi Alignment Algorithm.

**S:** 我正在這個中心從這個中心到這個公園要多久  
**P:** 

*I am now at this center. I need to go from this center to this park. How much time will it take?*

---

**Result** 我正在中國建築文化中心從中國建築文化中心到紫竹院公園要多久  
*I am now at the China Architectural Culture Center. I need to go from the China Architectural Culture Center to the Purple Bamboo Park. How much time will it take?*



<p><b><math>P_3</math>: stroke NUM=1</b>  <math>P_3[0]</math>: 紫竹院公園 -1  <math>P_3[1]</math>: 南長河 15.5                      ...</p>	<p><math>C_M(1,3)=1</math>  <math>C_T(1,3 1,2)=1</math>  <math>C_A(1,2)=1</math>  <math>C_A(1,3)=3</math></p>	<p><math>C_M(2,3)=1</math>  <math>C_T(2,3 1,2)=0</math>  <math>C_A(1,2)=1</math>  <math>C_A(2,3)=2</math></p>	<p><math>C_M(3,3)=0</math>  <math>C_T(3,3 2,2)=0</math>  <math>C_A(2,3)=0</math>  <math>C_A(3,3)=0</math>                      Optimal alignment</p>
<p><b><math>P_2</math>: stroke NUM=1</b>  <math>P_2[0]</math>: 中國建築文化中心 -1  <math>P_2[1]</math>: 三里河路 8                      ...</p>	<p><math>C_M(1,2)=0</math>  <math>C_T(1,2 1,1)=1</math>  <math>C_A(1,1)=0</math>  <math>C_A(1,2)=1</math></p>	<p><math>C_M(2,2)=0</math>  <math>C_T(2,2 1,1)=0</math>  <math>C_A(1,1)=0</math>  <math>C_A(2,2)=0</math></p>	<p><math>C_M(3,2)=1</math>  <math>C_T(3,2 2,1)=0</math>  <math>C_A(2,1)=1</math>  <math>C_A(1,3)=2</math></p>
<p><b><math>P_1</math>: point NUM=1</b>  <math>P_1[0]</math>: 中國建築文化中心 -1  <math>P_1[1]</math>: 三里河路 20                      ...</p>	<p><math>C_M(1,1)=0</math>  <math>C_A(1,1)=0</math></p>	<p><math>C_M(2,1)=0</math>  <math>C_T(2,1 1,1)=1</math>  <math>C_A(1,1)=0</math>  <math>C_A(2,1)=1</math></p>	<p><math>C_M(3,1)=1</math>  <math>C_T(3,1 2,1)=1</math>  <math>C_A(2,1)=1</math>  <math>C_A(3,1)=3</math></p>
<p><b>Generated hypotheses lists of speech and pen gestures</b></p>	<p><b><math>S_1</math>: NUM=1</b> 中國建築文化中心</p>	<p><b><math>S_2</math>: NUM=1</b> 中國建築文化中心</p>	<p><b><math>S_3</math>: NUM=1</b> 宋慶齡兒童科學公園 紫竹院公園</p>

Figure 5.2: An Illustrative Example on the Viterbi Alignment Algorithm. The arrows are the back pointer of  $(S_r, P_q)$ , which has minimum cumulative cost  $C_A(S_r, P_q)$ .

**Multimodal input**

S: 我 正在 “這個中心” 從 “這個中心” 到 “這個公園” 要 多久

*I am now at “this center”. I need to go from “this center” to “this park”.  
How much time will it take?*

P: 

**Unimodal paraphrase generated**

我 正在 “中國建築文化中心” 從 “中國建築文化中心” 到 “紫竹院公園” 要 多久

*I am now at “the China Architectural Culture Center”. I need to go from  
“the China Architectural Culture Center” to “the Purple Bamboo Park”.  
How much time will it take?*

Details will be described in Section 5.2. For an indirect SLR that does not have any corresponding aligned pen gesture, it will remain intact in the expression and will be further disambiguated through context inheritance in the dialog model of the spoken dialog system (SDS). An illustrative example is given in Table 5.2.

**5.1.4 Evaluating the Cross-Modality Integration Procedure**

We applied the cross-modality integration procedure to both the training and test sets. Recall that thus far we have been working with hand-transcribed speech input (with perfect SLR extraction performance), together with manually annotated gesture types for pen input. The transcriptions for speech and pen are regarded as perfect. For each multimodal inquiry, we manually annotate the alignment between an SLR and a pen gesture. Based on the alignment, the user’s intended location(s) can be identified. Similarly, the Viterbi alignment is applied to each multimodal inquiry so as to obtain a system generated alignment. A multimodal inquiry is considered as correct if the following two criteria are satisfied: (1) if the oracle (i.e. the manually annotated alignment)

**MM1:** *S* 從我“所在的地方”到“這裡”要多久？

*P.* • (*point to a hotel on the map*)

*How much time will it take to travel from “my current location” to “here”?*

**UM1:** 從“所在地”到“凱來大酒店”要多久？

*How much time will it take to travel from “my current location” to the “Gloria Hotel”?*

*Remarks: The system understands that “我所在的地方” is referring to the “current location,” which can be obtained from the dialog discourse. Also, 這裡 “here” can be jointly interpreted with the pointing gesture, due to high semantic compatibility (based on scoring). Therefore, UM1 contains the interpretations for both SLRs.*

**MM2:** *S:* 從“這個酒店”到“這個地方”有什麼車可以搭？

*P* → (*a stroke to indicate a street*)

*Which means of transportation can I use to travel from “this hotel” to “this place”?*

**UM2:** 從“這個酒店”到“王府井大街”有什麼車可以搭？

*Which means of transportation can I use to travel from “this hotel” to “Wangfujing Street?”*

*Remarks: The system can match 這個地方 “this place” with the stroke, due to high semantic compatibility (based on scoring). However, the indirect reference 這個酒店 “this hotel” cannot be matched with any pen gesture. Therefore, this SLR remains intact in UM2.*

Table 5.2: An example illustrating the unimodal paraphrases generated from the multimodal expressions from two dialog turns. MM1 and MM2 are the multimodal expressions from dialog turns 1 and 2. UM1 and UM2 are the unimodal paraphrases generated from MM1 and MM2 respectively. Translations are italicized.



and system generated alignments completely agree with each other; and (2) if the manually identified locations and the automatically identified intended locations completely match with each other. The cross-modality integration accuracy is defined as:

$$\text{Cross-modality integration accuracy} = \frac{G}{M} \quad (5.9)$$

where  $G$  is the total number of multimodal inquiries with perfect match between the oracle and system generated alignments and identified locations; and  $M$  is the total number of multimodal inquiries with SLRs in the (training or testing) data set.

The cross-modality integration procedure generated correct alignments between SLRs and pen gestures for 98.1% (950/968) of the training inquiries and 95.9% (416/434) of the testing inquiries that contain SLR(s). The incorrect pairings shed light on possible future work, including the need to use timing information across modalities for some multimodal inputs; as well as the need to apply pragmatic knowledge to infer the value of the NUM feature (i.e. in the case NUM= *nil*) and to filter out redundant SLRs in the speech input. An example with redundant SLRs is shown in Table 5.3. In the example, the user says 這個 (i.e. *this*) for four times to indicate four locations but he also mentions “這四個地方” (i.e. *these four places*) to confirm the number of locations he indicated. Incorporation of temporal information (i.e. temporal difference) may generate correct alignment.

<b>Reference transcription:</b>	
S:	我從“這裡”要到“這個”“這個”“這個”“這個”“這四個地方”一共需要多少時間
P:	• • • •
	<i>How much time will it take from “here” to “this”, “this”, “this” and “this”, “theses four places”?</i>
<b>Result of Cross-Modality Integration:</b>	
S:	我從“這裡”要到“這個”“這個”“這個”“這個”“這四個地方”一共需要多少時間
P:	• • • •

Table 5.3: An example on the incorrect alignment due to the presence of redundant SLRs (i.e. four “this” and one “these four places”) in the speech input.

## 5.2 Analytical Comparison between Parallel Multimodal and Unimodal Expressions

In order to investigate the relationships between speech and pen gestures and their effects in the joint interpretation, we performed an analytical comparison between collected multimodal expressions and their automatically generated unimodal paraphrases [63]. In order to do so, we ran the cross-modality integration procedure on the multimodal expressions. For each pair of aligned SLR and pen gesture, we can identify the user’s intended location(s). If we replace each of the SLRs with the full name of the identified location(s), we obtain the unimodal paraphrase. The correct paraphrases (over 98% of the entire data set) are extracted and combined with their semantically equivalent multimodal counterparts to form parallel corpora. More specifically, we obtain 984 multimodal and unimodal expression pairs from our training set and 422 pairs from our testing set. Comparative statistics of the multimodal and unimodal inputs from our training set are shown in Table 5.4. The total number of words are different due to the may due to the use of plural and ag-

gregated references in the multimodal inputs. For example, 這兩個購物中心 (i.e. *these two shopping centers*) will generate the full name of two shopping centers, 新東安市場 (i.e. *Xindong'an Plaza*) and 東方廣場 (i.e. *the Oriental Plaza*) in the unimodal paraphrases. We see that the spoken components of multimodal inputs are generally shorter and cover a smaller vocabulary than their unimodal counterparts. The difference is less pronounced than expected. One reason, based on our observation, is the diversity of spoken deictic expressions and Chinese measure words. For example, “*my current location*” may be verbalized in many ways (such as 身處點, 所在地, 目前所在的地方, 現在的地方, 現在這裡, 我的位置, 我的當前位置, 當前的位置, 我現在的地方, 我現在的地點, 我當前位置, etc.) Chinese measure words relating to location types (including 間, 個, 所, 條, 邊, 頭, 裡, 片, 帶, 塊, 點, 米, 圈, 塊兒, etc.) also contribute towards alternatives in verbalization.

	Multimodal input	Unimodal paraphrase
Total number of words	12,748	12,853
Average utterance length (in words)	8.8	8.9
(in chars.)	17.9	20.8
Range of utterance length (in words)	1 to 19	1 to 19
(in chars.)	2 to 54	2 to 58
Vocabulary size (number of words)	473	492

Table 5.4: Parallel multimodal and unimodal corpora statistics. The difference in the total number of words in multimodal input and unimodal paraphrase may be due to the use of plural and aggregated references in the multimodal input.

### 5.2.1 Language Modeling

We pooled the multimodal and unimodal spoken expressions together (1,450 in all as presented in [56]) to train a class trigram language model. We classified the proper names (i.e. location names) into 12 equivalences classes, e.g. UNIVERSITY, HOSPITAL, STREET, etc. We also have 4 other equivalences classes including: ARTICLES, NUMBERS (i.e. implicit/explicit numeric expressions, e.g. — *one*, 幾 *few*, 些 *some*, etc.), MEASURE\_WORDS and LOCATION\_TYPE (e.g. the words 大學 *university*, 公園 *parks*, etc.) The language model was developed using the CMU SLM toolkit [64]. The resulting model contains 290 unigrams, 1,375 bigrams and 2,795 trigrams. The probabilities are smoothed by Katz backoff smoothing [65] with discount ratios 0.04 for unigrams, 0.36 for bigrams, and 0.38 for trigrams. The discounting thresholds for unigrams, bigrams and trigrams are 1, 5 and 7 respectively. The discount ratios and discounting thresholds are determined by the CMU SLM toolkit automatically using the inquiries from the training set. We computed the class trigram perplexities for the multimodal and unimodal test sets respectively. Results are shown in Table 5.5.

Comparisons in Class Trigram Test Set Perplexities		
	Multimodal Input	Unimodal Paraphrases
Total number of utterances	422	422
Number of words	4,505	4,555
Perplexity ( $PP$ )	16.5	29.5

Table 5.5: Comparisons in perplexities between the parallel multimodal (MM) and unimodal (UM) inputs. The difference in the number of words is less than expected due to the diversity of Chinese measure words and contextual phrases mentioned.

We observe that for the semantically equivalent, parallel multimodal and

unimodal corpora, the unimodal paraphrases have significantly higher perplexities. We also observe from Table 5.6 that the test set may be divided into two subsets according to comparisons in per-utterance perplexities between the multimodal ( $PP_{MM}$ ) and unimodal inputs ( $PP_{UM}$ ) for further analysis.

Comparisons in Per-Utterance Perplexities	
$PP_{MM} < PP_{UM}$	349/422 utterances (82.7%)
$PP_{MM} = PP_{UM}$	73/422 utterances (17.3%)
$PP_{MM} > PP_{UM}$	0

Table 5.6 Comparison of per-utterance perplexities between the multimodal inputs and their unimodal paraphrases

### 5.2.2 Data Analysis

#### (A) Category ( $PP_{MM} = PP_{UM}$ )

As shown in Table 5.6, the testing data subset with this inequality contain 17.3% (73/422) of the expressions. For this category, we found that the majority (66%, 48/73) of the expressions involve *redundancy* between the speech and pen modalities. Redundancy means “the same piece of information/semantic content is carried by both modalities.” As shown in Example 1 of Table 5.7, each pair of (x,y) coordinates of each pointing gesture in the multimodal input *matches with* the abbreviation of the location name that was uttered. The unimodal paraphrase incorporates the full name of each location during generation. However, since our class-based language model gives the same probability values to both the abbreviated and full names of the same location, the per-utterance perplexity values are the same.

Example 2 in Table 5.7 illustrates the use of ellipsis, which occurred for (33%, 24/73) of the cases in this testing data subset. The subject inputs four

pen strokes that connects four locations and simply uttered 最快的交通路線 “the fastest route”. We interpret that the subject wishes to obtain the fastest route that traverses the four indicated locations. However, the speech modality does not mention the locations at all. Hence the cross-modality integration framework cannot capture the ellipsis and generate a unimodal paraphrase that ignores the pen gestures, resulting in an equal perplexity value. This is an artifact because in reality the multimodal expression conveys a greater amount of information when compared to its unimodal paraphrases. Ellipsis should be a case of complementarity across modalities where certain semantic content appears in one modality and is completely omitted from the other modality.

Example 3 illustrates the occurrence of a spoken locative reference expression that is redundant with the pointing gesture, followed by an ellipsis. Again, we observe equal per-utterance perplexities and the explanations are consistent with the two previous examples. Redundancy between the speech and pen modalities should be very useful in face of imperfect recognition outputs, e.g., in automatic speech and pen gesture recognitions. Handling ellipsis merits further investigation for automatic interpretation of multimodal input. A possible method to handle ellipsis is to integrate the semantics from pen gesture to the recognized speech input according to the time of occurrence.

**(B) Category ( $PP_{MM} < PP_{UM}$ ):**

There are 422 expressions in the test set, of which, the testing data subset with this inequality contains 349 (82.7%, 349/422) expressions. Expressions in this category involve *complementarity* between the speech and pen modalities. Complementarity means “a piece of information/semantic content is carried across multiple modalities, i.e. either modality alone is semantically ambiguous and a clear semantic meaning can be obtained when semantics across

<p><b>Example 1:</b></p> <p><b>Multimodal Expression, <math>PP_{MM} = 3.61</math></b></p> <p>(Note redundancy across modalities)</p> <p>S 從“北郵”到“北航”“地質大學”“北科大”和“北醫”要多久</p> <p><i>How much time will it take from “BUPT” to “Beihang”, “CUG”, “USTB” and “BJMU”?</i></p> <p>P        •        •        •        •        •</p> <p><b>Unimodal Paraphrase, <math>PP_{UM} = 3.61</math></b></p> <p>從“北京郵電大學”到“北京航空航天大學”“中國地質大學”“北京科技大學”和“北京醫科大學”要多久</p>
<p><b>Example 2:</b></p> <p><b>Multimodal Expression, <math>PP_{MM} = 4.93</math></b></p> <p>(Note ellipsis)</p> <p>S            最快的交通路線 <i>The fastest route.</i></p> <p>P. — — — →</p> <p><b>Unimodal Paraphrase, <math>PP_{UM} = 4.93</math></b></p> <p>最快的交通路線 <i>The fastest route.</i></p>
<p><b>Example 3:</b></p> <p><b>MM Expression, <math>PP_{MM} = 654.3</math></b></p> <p>S “我的位置”        交通路線 <i>“my current location”. Travel route please.</i></p> <p>P.        •        →</p> <p><b>UM Paraphrase, <math>PP_{UM} = 654.3</math></b></p> <p>“身處點” 交通路線 <i>“my current location”. Travel route please.</i></p>

Table 5.7: Illustrative examples from the testing data subset with ( $PP_{MM} = PP_{UM}$ ).

modalities are combined.” We present illustrative examples in Table 5.8. As shown in Example 1, the speech and pen modalities complement each other in specifying a group of intended locations. Either modality alone is semantically ambiguous, e.g., the spoken expression “*here*” that corresponds to the point, or the expression “*these universities*” that correspond to the circle. However, when the semantics across modalities are combined, the semantic meaning is clear. Hence we can see that part of intended message is conveyed via the speech modality, while the remaining part is conveyed via the pen modality. The unimodal paraphrase, however, captures the full semantics of the subject’s intended message. Consequently, the perplexity of the spoken component in the multimodal expression is less than that of the unimodal paraphrase.

Example 2 in Table 5.8 illustrates the possibility that a multimodal expression can exhibit both redundancy and complementarity in sequential locative reference expressions. The first rendition shows five reference expressions, all of which exhibit complementarity between the speech and pen modalities. Among the 349 expressions in this data subset involve complementarity, there are 321 (92%, 321/349) similar cases (i.e. complementarity across modalities) in this data subset. The second rendition shows redundancy in the first reference expression, while the remaining four expressions exhibiting complementarity. Hence the per-utterance perplexity rose slightly (c.f. the first rendition) even though both renditions are semantically equivalent. There are 28 (8%, 28/349) similar cases (i.e. combined redundancy and complementarity) in this data subset. The third rendition is the unimodal paraphrase, which has the highest per-utterance perplexity value. Table 5.9 shows the overall statistics of the categories.

The example in Table 5.10 also illustrates the advantage of perplexity reduction by virtue of complementarity across the speech and pen modalities, through comparison between the speech components in a multimodal expres-



**Example 1:****Multimodal Expression,  $PP_{MM} = 4.53$** 

(Note complementarity across modalities)

S: 我現在在“這裡”我想分別去“這幾所大學”有哪些交通線路可以選擇

*I am now “here”. I want to visit “these universities”. What are the possible travel routes?*

P:           •                           ○

**Unimodal Paraphrase,  $PP_{UM} = 6.50$** 

我現在在“北京電影學院”我想分別去“北京航空航天大學”“北京科技大學”“中國地質大學”“北京醫科大學”有哪些交通線路可以選擇

*I am now at “Beijing Film Academy”. I want to visit “Beihang University”, “China University of Geosciences”, “University of Science and Technology Beijing” and “Beijing Medical University”. What are the possible travel routes?***Example 2:****First rendition - Multimodal Expression,  $PP_{MM} = 5.71$** 

(Note complementarity across modalities)

S: 從“這裡”到“這裡”“這裡”“這裡”還有“這裡”有什麼交通路線

*What is the travel route from “here” to “here”, “here”, “here” and “here”?*

P:           •           •           •           •           •

**Second rendition - Multimodal Expression,  $PP_{MM} = 9.08$** 

(Note redundancy in the first reference expression and complementarity in the remaining four expressions)

S: 從“北郵”到“這裡”“這裡”“這裡”還有“這裡”有什麼交通路線

P:           •           •           •           •           •

**Unimodal paraphrase  $PP_{UM} = 9.21$** 

從“北京郵電大學”到“北京航空航天大學”“北京科技大學”“中國地質大學”還有“北京醫科大學”有什麼交通路線

Table 5.8: Illustrative examples from the testing data subset with ( $PP_{MM} = PP_{UM}$ ).

Total number of expressions		422
$PP_{MM} < PP_{UM}$ (349 expressions)	Complementarity	321/349
	Complementarity and Redundancy	28/349
$PP_{MM} = PP_{UM}$ (73 expressions)	Redundancy	48/73
	Ellipsis	24/73
	Redundancy and Ellipsis	1/73
$PP_{MM} > PP_{UM}$ (0 expression)	-	-

Table 5.9: Overall statistics of different categories found by comparison of per-utterance perplexities between the multimodal inputs and their unimodal paraphrases.

sion with its counterpart in a unimodal expression. In particular, the unimodal expression in Example 1 in Table 5.10 has a perplexity of 25.1, which is reduced to 5.9 in a multimodal expression (see Example 2) with complementary speech and pen inputs. However, if the speech and pen inputs are redundant, as shown in Example 3, there is no perplexity reduction. If there is a mixture of complementary and redundant inputs between the two modalities (see Example 4), then there is a smaller reduction in perplexity from 25.1 to 8.8.

### (C) Findings and Implications

Categorization of the test set based on perplexity values, followed by analysis of the categories enables us to visualize the effects of complementarity and redundancy [43] across the speech and pen modalities in multimodal user inputs.

Complementarity offers expressive power, because the user is free to distribute various parts of the message to different modalities to ease (complex)

<p><b>Example 1 - <math>PP_{UM} = 25.1</math> (generated unimodal paraphrase)</b></p> <p>從“北京郵電大學”到“北京航空航天大學”“中國地質大學”“北京醫科大學”和“北京科技大學”要多久</p> <p><i>How much time will it take from “Beijing University of Posts and Telecommunications” to “Beihang University”, “China University of Geosciences”, “Beijing Medical University” and “Beijing University of Science and Technology”?</i></p>
<p><b>Example 2 - <math>PP_{MM} = 5.9</math> (complementarity)</b></p> <p>S: 從“這裡”到“這四所大學”要多久</p> <p>P:       •       ••••</p> <p><i>How much time will it take from “here” to “these four universities”?</i></p>
<p><b>Example 3 - <math>PP_{MM} = 25.1</math> (redundancy)</b></p> <p>S. 從“北郵”到“北航”“地大”“北醫”和“北科”要多久</p> <p>P:       •       •       •       •       •</p> <p><i>How much time will it take from “BUPT” to “BUAA”, “CUG”, “BMU” and “BUST”?</i></p>
<p><b>Example 4 - <math>PP_{MM} = 8.8</math> (complementarity and redundancy)</b></p> <p>S. 從“北郵”到“這四所大學”要多久</p> <p>P:       •       ••••</p> <p><i>How much time will it take from “BUPT” to “these four universities”?</i></p>

Table 5.10: Examples illustrating perplexity reduction in different cases. Translations are italicized.

communication and to reduce cognitive loading [42]. Semantic decoding of an individual modality generates a partial interpretation of the intended message and these partial semantics need to be integrated in order to gain a complete understanding of the user's intent. This motivates us to use a late semantic fusion architecture for multimodal input interpretation.

Redundancy occurs when both the speech and pen modalities carry the same semantic content. As a preliminary step, the current work only deals with perfect transcriptions of the speech recordings and filtered pen gesture recognition outputs. However, we may conceive that in real applications, the recognition outputs (that will be presented in Section 5.3) corresponding to different input modalities may be erroneous. Redundancy across modalities motivates the use of mutual disambiguation techniques [66]. In addition, we also observe occurrences of ellipses, where some locative references are omitted from the speech component in the multimodal expression and is expressed only with the pen component. Ellipses motivate further investigations in the syntax of the multimodal language, as well as the use of such multimodal integration approaches as finite-state transducers [50].

### 5.3 Cross-Modality Integration on Imperfect Transcriptions

The cross-modality integration procedure has demonstrated reasonable performance in perfect transcriptions in Section 5.1, which is acted as the upper bound of the integration performance. However, under practical situations, captured inputs are much more problematic, due to disfluencies in the speech modality (e.g. filled pauses and repairs), spurious pen gestures and recognition errors in both modalities. These imperfections have adverse effects on cross-modality integration. Therefore, in this section, we attempt to apply

the cross-modality integration procedure on imperfect transcriptions.

### 5.3.1 Transcribing the Spoken Inputs

We transcribed the speech signals in the multimodal corpus with a Mandarin speech recognizer [67] that is developed with the HTK toolkit [68]. This recognizer was originally trained with 75 hours of read-speech recorded in a clean environment from a general open domain (i.e. newspaper). Hence, we replaced the recognizer’s general-domain lexicon with a domain-specific version of 637 entries that contain names of locations in Beijing as well as frequent spoken deictic expressions. We also incorporated a domain-specific bigram language model trained from manual transcripts of the training data set. The acoustic models remain unchanged. Speech recognition performance evaluated based on the top-scoring recognition hypotheses gave overall character accuracy of 44.6%. In particular, we observe that performance is especially poor for four of the subjects who spoke Mandarin with an accent. Further degradation was due to background noise. Speech recognition performance evaluated based on the top-scoring recognition hypotheses across subjects are shown in Figure 5.3. Application of the SLR extraction procedure (see Section 4.1) to the top-scoring recognition hypotheses shows substitution, deletion and insertion errors in the SLRs. SLR deletion and substitution are the most prominent, frequently caused by short duration of 這兒 (meaning *here* and pronounced as /zher/) and phonetic confusion between 這 (meaning *this* and pronounced as /zhe/) and 車 (meaning *car* and pronounced as /che/). Another example of phonetic confusion is between the second character 裡 (meaning *inside* and pronounced as /li/) of 這裡 (meaning *here* and pronounced as /zheli/) and 米 (meaning *rice* and pronounced as /mi/).

For each spoken input expression, we compare the list of parsed SLR(s) with its oracle transcription with the list of parsed SLR(s) from its speech

recognition transcription in order to check the SLR recognition performance. The SLR recognition accuracy is defined as:

$$SLR \text{ Recognition Accuracy} = \frac{N_{SLR} - I_{SLR} - S_{SLR} - D_{SLR}}{N_{SLR}} \quad (5.10)$$

where  $N_{SLR}$  is the total number of SLRs in the oracle transcriptions;

$I_{SLR}$ ,  $S_{SLR}$  and  $D_{SLR}$  are the numbers of insertion, substitution and deletion errors from the speech recognition transcriptions respectively.

Overall, the SLR recognition accuracies (each SLR is treated as a word) for the training and test sets are 38.5% and 39.3% respectively. Furthermore, only 55.6% of the direct references and 29.1% of the indirect references can be recognized correctly in the training. In other words, over half of the SLRs have not been correctly extracted. However, the majority (>60%) of the incorrectly recognized SLRs involves confusion with other SLRs carrying the same semantic meaning. In this work, the confusion between SLRs during speech recognition may involve only the measure word and abbreviation and hence does not alter the semantic meaning. For example, both 這個大學 and 這大學 mean *this university*; and both 所在的地方 and 所在地 mean *current location*. Hence, these incorrectly recognized SLRs will not affect the subsequent cross-modality integration process. Overall, 50.9% and 51.7% of the recognized SLR in training and test sets were interpreted with correct semantics.

### 5.3.2 Recognizing the Pen Inputs

We have developed a pen gesture recognizer, based on a simple algorithm that proceeds through a sequential procedure of recognizing a point, a circle and a stroke, as follows.

- (i) **Recognizing Points:** If the pixel distance between the pen down and pen up coordinates is fewer than  $q$  ( $= 6$ ) pixels, which is the width of a square

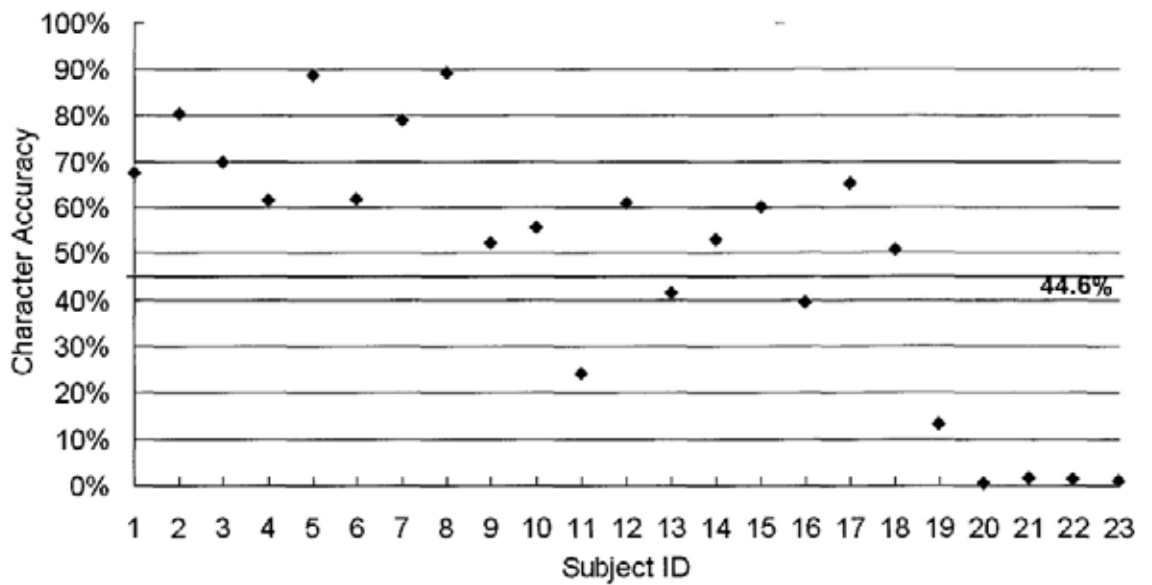


Figure 5.3: Speech recognition performance (character accuracies) across subjects in the training set of the multimodal corpus.

icon on the map, the input is considered as a point. Detected pointing actions with temporal difference less than 0.25 second are considered repetitive and the redundancy will be discarded. If the pen gesture is not classified as a point, it will be evaluated as a circle or stroke, described as follows.

- (ii) **Recognizing Circles:** A pen gesture is recognized as a circle if 80% of its  $x$  and  $y$  coordinates appears at least twice and if it contains no more than two convex hulls (Figure 5.4 shows an example of convex hull). If the pen gesture is not classified as a circle, it will be evaluated as a stroke, described as follows.
- (iii) **Recognizing Strokes:** Since strokes are directional, a pen gesture is recognized as stroke if one or both of the  $x$  and  $y$  coordinates shows directional migration towards pen down coordinates, also the radius of



## A convex hull

Figure 5.4: An illustration of convex hull in a circle.

curvature (ROC) cannot be below a preset threshold of 24 [23]. The ROC of a circle formed by three points  $x$ ,  $y$  and  $z$  as shown in Figure 5.5 is given in Equation 5.11.<sup>12</sup> If the pen gesture is not classified as a stroke, it will be rejected.

$$ROC = r(x, y, z) = \frac{|x - y| |y - z| |z - x|}{A(x, y, z)} \quad (5.11)$$

where  $A(x, y, z)$  is the area of triangle formed by  $x$ ,  $y$  and  $z$  and

$|\cdot|$  denotes the Euclidean distance between the two coordinates.

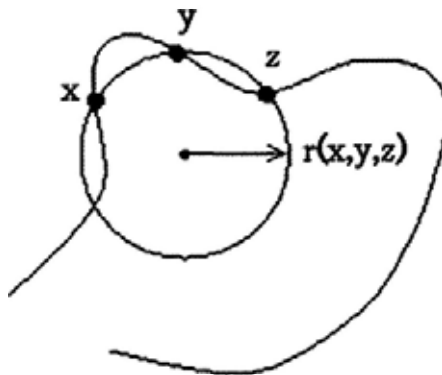


Figure 5.5: A circle formed by three points,  $x$ ,  $y$  and  $z$  and the radius of curvature.

This simple pen gesture recognition algorithm can only generate a single output hypothesis, which will be the top best pen gesture recognition output. Overall pen gesture recognition accuracy is 86.6%. Table 5.11 shows some pen

<sup>12</sup>Figure 5.5 is borrowed from [69]



gesture recognition errors. The incorrectly recognized pen gestures include confusions that may carry the same semantic meaning and hence the pen recognition error will not affect the subsequent integration process. Overall, 91.3% of the recognized pen gestures can be interpreted with correct semantic meaning.



A flat circle is mis-recognized as STROKE.	
A distorted stroke with low ROC, which is rejected by the recognizer.	

Table 5.11: Illustrative examples on the recognition errors of circle and stroke.

### 5.3.3 Evaluating the Cross-Modality Integration Procedure

We applied the cross-modality integration procedure to each multimodal inquiry of both training and test sets of imperfect speech and pen transcriptions (which contain speech and pen gesture recognition errors and repetitive pen gesture inputs) so as to obtain a system generated alignment. Comparison between the system generated alignments with the manually annotated alignments shows that the cross-modality integration procedure generated correct alignments between SLRs and pen gestures for 51.1% (495/968) of the training inquiries and 54.4% (236/434) of the testing inquiries. Performance statistics of the cross-modality integration procedure are shown in Table 5.12 and Figure 5.6. The performance achieved is better than expected at a speech recognition accuracy of 44.6%.<sup>13</sup> Analysis shows that this is because of the complementarity relation between speech and pen modalities, where the two modalities mutually disambiguate [70] with each other in the presence of recognition

<sup>13</sup>If we assume that the two modalities are independent of each other, the expected performance is  $(\text{Accuracy}_{\text{Speech Recognition}} \times \text{Accuracy}_{\text{Pen Recognition}}) = 44.6\% \times 86.6\% = 38.6\%$ .

errors. Integration of the hypothesis lists generated from the mis-recognized SLR and pen gestures shows that the cross-modality integration procedure can still extract the common location(s) found, which may be correct. Table 5.13 shows an illustrative example on mutual disambiguation between the two modalities in cross-modality integration with the presence of recognition errors.

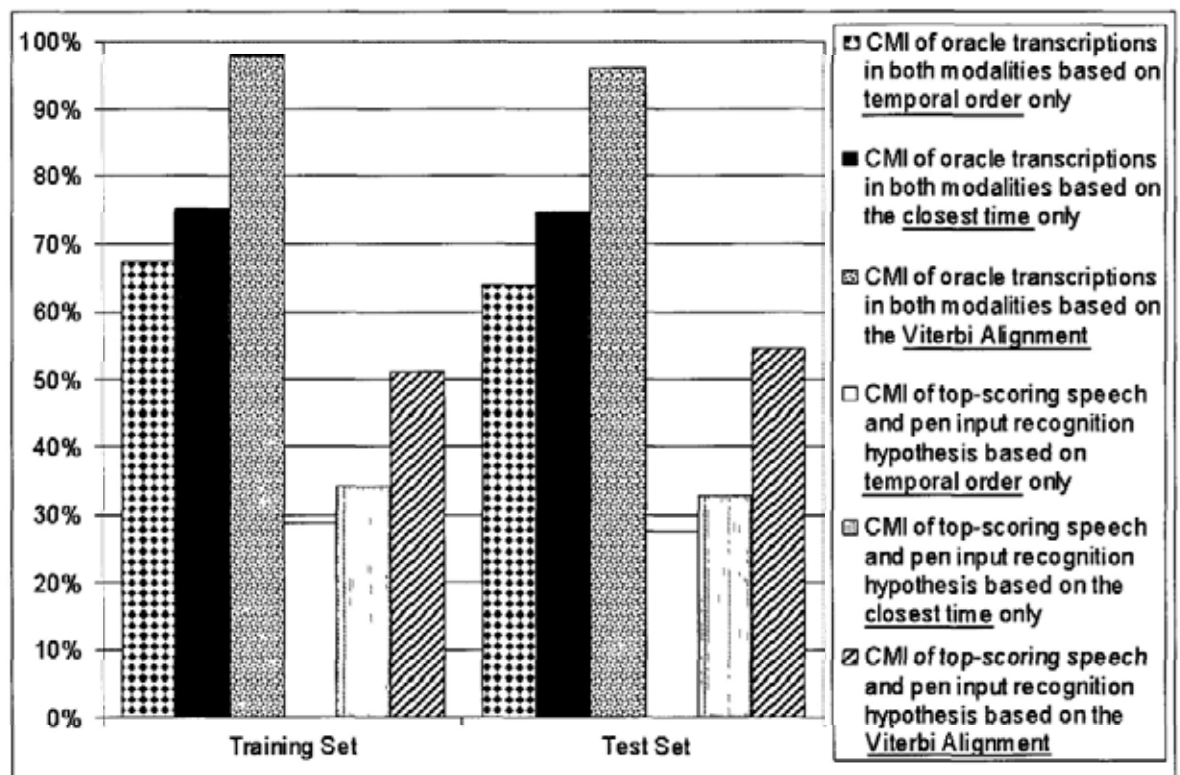


Figure 5.6: Performance of the cross-modality integration (CMI) in the training and test sets.

We also applied the cross-modality integration procedure to each multi-modal inquiry in both training and test sets, according to manual (which is considered as perfect) and imperfect speech and pen transcriptions (which contain errors). In this way, we can analyze the reliance of the cross-modality integration procedure on speech and pen gestures recognition accuracies. Per-

formance statistics of the cross-modality integration procedure are shown in Table 5.14 and Figure 5.7. Since the overall pen gesture recognition accuracy is relatively high (i.e. 86.6% of the pen gesture inputs can be recognized correctly), the performance difference between cross-modality integration with oracle transcriptions (i.e. perfect) and recognition hypothesis (i.e. imperfect) of pen inputs is small (i.e. comparison between the third and fourth rows; and the last two rows of Table 5.14). However, since the overall speech recognition accuracy is relatively low (i.e. speech recognition character accuracy of 44.6%), the performance difference between cross-modality integration with oracle transcriptions (i.e. perfect) and recognition hypothesis (i.e. imperfect) of speech input is larger (i.e. comparison between the third and fifth rows; and the fourth and the last rows of Table 5.14). If we assume that there is linear correlation between the performance of cross-modality integration and overall speech recognition accuracy, the goal of overall speech recognition accuracy need to be 77.3% and 90% so as to achieve an cross-modality integration performance of 80% and 90% respectively (as shown in Figure 5.8).

## 5.4 Chapter Summary

In this chapter, we have described our work in semantic integration of multimodal user inputs that consist of speech and pen gestures. Partial interpretations from individual modalities are combined using Viterbi alignment, which enforces the constraints of temporal order and semantic compatibility constraints in its cost functions to generate an integrated interpretation across modalities for overall input. Experiments show that this approach can correctly interpret around 98% and 96% of the multimodal inquiries in our training and test sets respectively. We have also performed a comparative analysis of multimodal (*MM*) user inputs together with their semantically equivalent unimodal (*UM*) counterparts. These are generated by the cross-

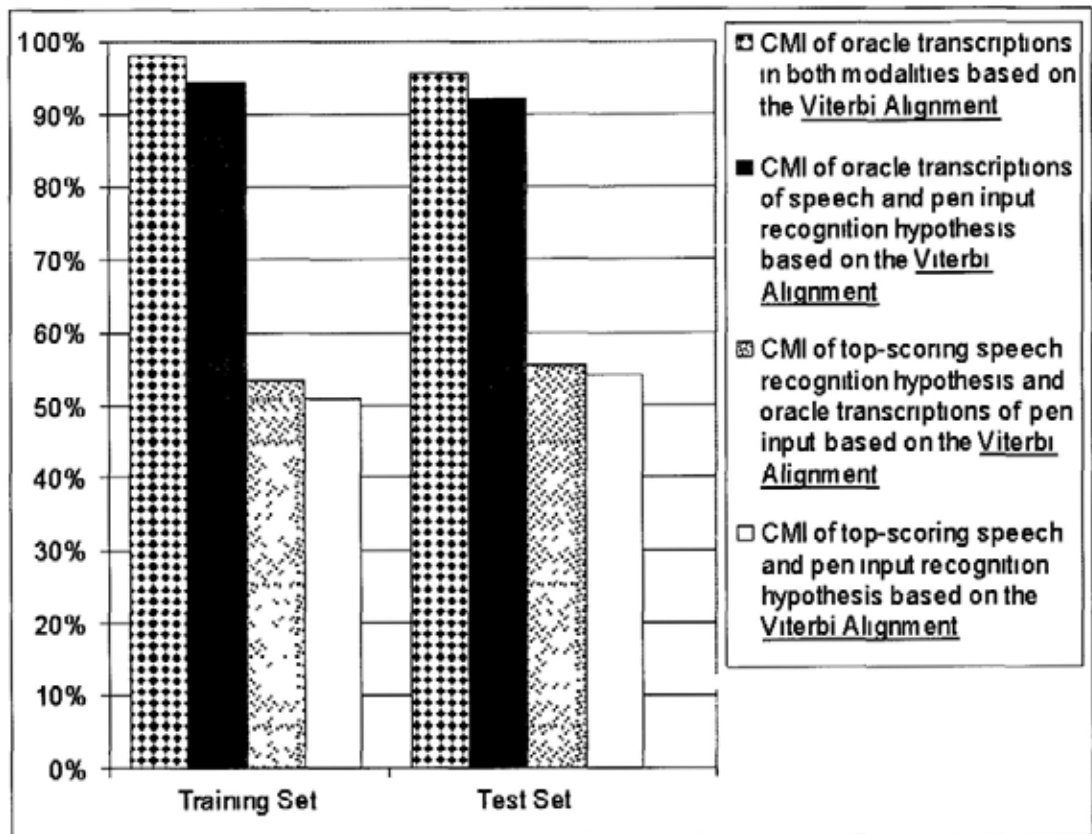


Figure 5.7 Performance of the cross-modality integration (CMI) in the training and test sets

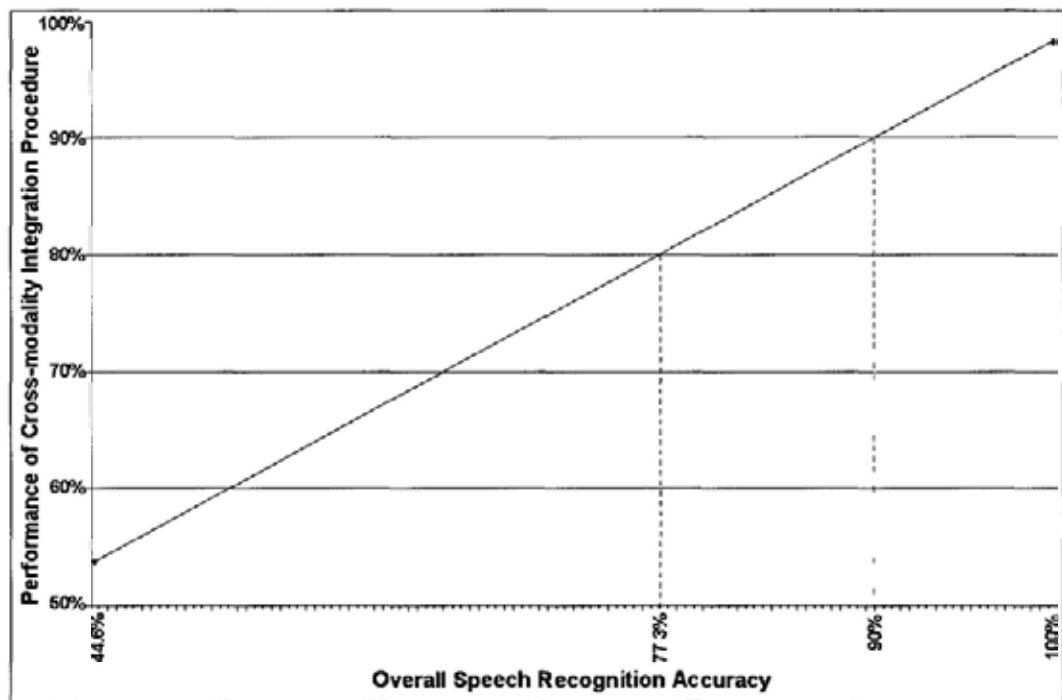


Figure 5.8: Plot of the relation between the performance of cross-modality integration procedure and overall speech recognition accuracy.

modality framework proposed. We trained a class trigram language model with 1,450 multimodal and unimodal speech utterances and compared the perplexities ( $PP$ ) between parallel multimodal and unimodal test sets (with 422 utterances each). We observe that the speech components of multimodal expressions are generally shorter with lower lexical variability than their unimodal counterparts. Comparison with per-utterance perplexities affirms the relationships of complementarity and redundancy across the speech and pen modalities. One subset of our data exhibits the equality of ( $PP_{MM} = PP_{UM}$ ) and consists mainly of multimodal expressions where speech and pen modalities carry redundant semantics. The other subset exhibits the inequality of ( $PP_{MM} < PP_{UM}$ ) where the speech and pen modalities carry complementary semantics. We also observe the occurrences of ellipsis, where certain semantics appear in one modality but not the other, and forms a special case of complementarity. These observations have implications on the choice of fusion architectures for multimodal input interpretation. In practical situation, speech and pen gesture inputs contain recognition errors and spurious inputs. Our Mandarin speech recognizer has an overall character accuracy of 44.6% and the pen gesture recognizer has an overall gesture type recognition accuracy of 86.6%. Application of the cross-modality integration framework on the imperfect recognition outputs shows that the proposed framework can correctly generate alignments between SLRs and pen gestures for around 51% and 54% of the multimodal inquiries in the training and test sets respectively. Analysis shows that complementarity relation between SLRs and pen gestures can salvage the performance of cross-modality integration in the present of recognition errors.

	Training Set	Test Set
Number of multimodal inquiries	1002	440
Number of multimodal inquiries that contain SLR(s)	968	434
Cross-modality integration of oracle transcriptions in both modalities based on <u>temporal order only</u> (i.e. align one-by-one)	67.3% (651/968)	64.1% (278/434)
Cross-modality integration of oracle transcriptions in both modalities based on <u>the closest time only</u> (i.e. simultaneous or smallest time lag)	75.1% (727/968)	74.4% (323/434)
Cross-modality integration of oracle transcriptions in both modalities based on the Viterbi Alignment in Chapter 5	98.1% (950/968)	95.9% (416/434)
Cross-modality integration of top-scoring speech and pen input recognition hypothesis based on <u>temporal order only</u>	28.8% (279/968)	27.4% (119/434)
Cross-modality integration of top-scoring speech and pen input recognition hypothesis based on <u>the closest time only</u> (i.e. simultaneous or smallest time lag)	34.3% (332/968)	32.9% (143/434)
Cross-modality integration of top-scoring speech recognition hypothesis and recognized pen inputs based on the Viterbi Alignment in Chapter 5 and in the Table 5.1 in Section 5.1	51.1% (495/968)	54.4% (236/434)

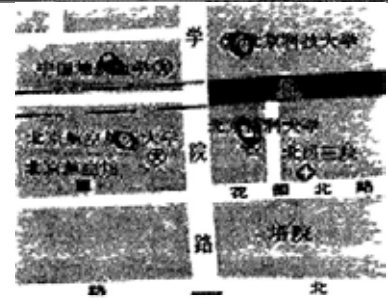
Table 5.12: Performance of the cross-modality integration, measured in terms of percentage of correctly aligned expressions in the training and test sets.

**Reference transcriptions:**

S: 從“我所在的地點”到“這四個大學”要多久？

P: ○○○○  
(four repeated circles)

How much time will it take from “my current location” to “these four universities?”

**Top-scoring speech and pen recognition hypotheses:**

S: 從和“現在的地方”到“這幾個地方”要多久？

P: ○●●○ (two repeated circles and two points)

How much time will it take from “here” to “these locations”?

**Hypothesis lists of recognized speech input:**

SLR1: EXISTING\_LOCATION=現在的地方 *current location*

所在地 *my current location*

SLR2: DEICTIC=這幾個地方 *these locations*

NUM=plural

LOC\_TYPE=nil

北京科技大學 *University of Science and Technology Beijing*

中國地質大學 *China University of Geosciences*

學院路 *Xueyuan Road*

北京航空航天大學 *Beihang University*

北京醫科大學 *Beijing Medical University*

北京航空館 *Beijing Aviation Museum*

北醫三院 *Peking University Third Hospital*

花園北路 *Huayuan North Road . . . . .*

(all locations on the map shown)



<p><b>Hypothesis lists of recognized pen input:</b> (<i>locations ranked by distance in pixels</i>)</p> <p>Pen1: TYPE=CIRCLE</p> <p>北京航空航天大學 <i>Beihang University</i></p> <p>北京航空館 <i>Beijing Aviation Museum</i> . . . . .</p> <p>Pen2: TYPE=POINT</p> <p>中國地質大學 <i>China University of Geosciences</i></p> <p>學院路 <i>Xueyuan Road</i> . . . . .</p> <p>Pen3: TYPE=POINT</p> <p>北京科技大學 <i>University of Science and Technology Beijing</i></p> <p>學院路 <i>Xueyuan Road</i> . . . . .</p> <p>Pen4: TYPE=CIRCLE</p> <p>北京醫科大學 <i>Beijing Medical University</i></p> <p>北醫三院 <i>Peking University Third Hospital</i> . . . . .</p>
<p><b>Generated unimodal paraphrase:</b></p> <p>從 和 “所在地” 到 “北京航空航天大學” “中國地質大學” “北京科技大學” “北京醫科大學” 要多久？</p> <p><i>How much time will it take from “my current location” to “Beihang University”, “China University of Geosciences”, “University of Science and Technology Beijing”, “Beijing Medical University”?</i></p>
<p><i>Remark: Mis-recognition of SLR1 does not change its semantic meaning and does not affect the integration process. The lost of LOC_TYPE feature of SLR2 leads to generation of a longer hypothesized list of locations. The alignment between four pen gestures and SLR2 can compensate the lost of NUM feature of SLR2. Due to the complementarity between speech and pen modalities, we can generate correct unimodal paraphrase.</i></p>

Table 5.13: Examples on the correct integration with the present of SLR recognition error.

	Training Set	Test Set
Number of multimodal inquiries	1002	440
Number of multimodal inquiries that contain SLR(s)	968	434
Cross-modality integration of oracle transcriptions in both modalities	98.1% (950/968)	95.9% (416/434)
Cross-modality integration of oracle transcriptions of speech and recognized pen inputs	94.5% (915/968)	92.4% (401/434)
Cross-modality integration of top-scoring speech recognition hypothesis and oracle transcriptions of pen inputs	53.7% (520/968)	55.8% (242/434)
Cross-modality integration of top-scoring speech recognition hypothesis and recognized pen inputs	52% (506/973)	52.4% (225/429)

Table 5.14: Performance of cross-modality integration, measured in terms of the percentage of correctly aligned expressions in the training and test sets based on the Viterbi Alignment in Chapter 5 and Table 5.1 in Section 5.1.

## Chapter 6

# Hypothesis Rescoring for Robustness towards Imperfect Transcriptions

The cross-modality integration procedure has demonstrated reasonable performance (around 97% accuracy) in aligning spoken locative reference (SLR) expressions with pen gestures in oracle-transcribed multimodal inputs. These transcriptions are essentially perfect. However, the performance drops to around 50% under practical situations with spurious pen gestures and recognition errors in both modalities. These imperfections have adverse effects on cross-modality integration. Therefore, in this chapter, we describe our attempt to extend the cross-modality integration procedure with the use of multiple recognition hypotheses in order to achieve robustness towards recognition errors. Consider the scenario in which a speech recognizer generates  $N$ -best hypotheses based on the speech input, while the pen gesture recognizer generates  $M$ -best hypotheses based on the pen input. The hypotheses are rank ordered according to their recognition scores in each individual modality. As such, we will have  $N \times M$  possible candidates for cross-modality integration.

In designing a rescoring mechanism for comparing these candidates for integration, we should consider such elements as the quality of the recognized spoken locative references, the quality of the interpreted pen gestures and the quality of the alignment. Figure 6.1 shows the system architecture of the extended cross-modality integration framework (i.e. cross-modal integration with hypothesis rescoring). We will elaborate on these points in the following subsections.

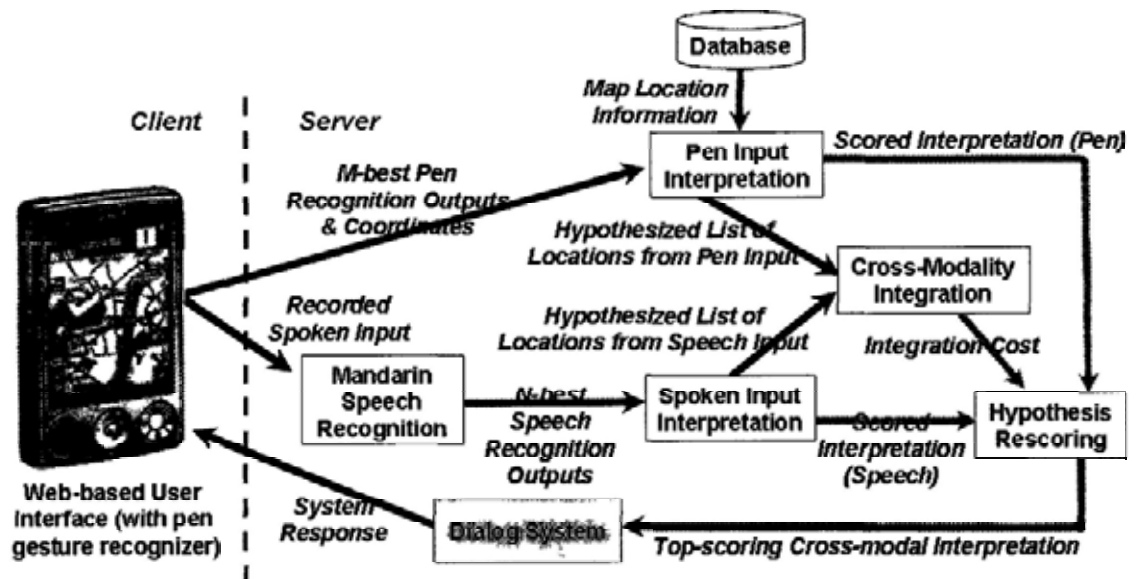


Figure 6.1: The system architecture of cross-modality integration with hypothesis rescoring, which can be the front-end multimodal processing framework for an existing unimodal dialog system.

## 6.1 Pruning and Scoring the Recognized Spoken Inputs

The speech recognizer can generate an  $N$ -best hypothesis list of speech recognition transcripts. However, the recognizer may generate non-sensical hypotheses in the  $N$ -best hypothesis list. We devise a pruning strategy based on perplexity to filter out non-sensical transcripts. A recognition transcript

with a small value of perplexity is more likely to have a reasonable interpretation. This is because the hypothesized word sequence generally conforms to the predictions by the  $n$ -gram language model. Hence our pruning strategy targets the opposite cases-hypotheses with large perplexity values exceeding a preset threshold are filtered.

The speech component of a multimodal input expression may be transcribed by speech recognition as a hypothesized word sequence with  $R$  spoken locative references (SLRs). For a segment of the speech signal with specific start and end times, we may observe transcriptions across the  $N$ -best ( $N = 100$  in this work) speech recognition hypotheses. Let  $S_r$  denote the  $r^{\text{th}}$  SLR in one of the speech recognition hypotheses, which is also the transcription of a specific speech signal segment. We may score the quality of this transcription by defining the normalized cost  $C_S(S_r[N])$  for the recognized SLR ( $S_r$ ), as shown in Equation 6.1.  $n(S_r[N])$  is the number of times the speech segment is transcribed as  $S_r$  across the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $\frac{n(S_r[N])}{N}$  is known as the  $N$ -best *purity* of the SLR  $S_r$ , where purity values range between zero and one. The higher the purity, the more preferable the SLR  $S_r$ , and the lower is the normalized cost of the speech transcription  $C_S(S_r[N])$ .<sup>1</sup> Table 6.1 presents an illustrative example on the normalized cost  $C_S(S_r[N])$  for recognized SLR  $S_r$  while the complete list of speech recognition hypotheses is shown in Table D.1 in Appendix D.

$$C_S(S_r[N]) = 1 - \frac{n(S_r[N])}{N} \tag{6.1}$$

---

<sup>1</sup>It is conceivable that should a pen gesture recognizer be used to generate  $M$ -best recognition hypotheses, a similar  $M$ -best purity may be incorporated in the cost function for the pen modality.

Reference transcription	
從“這兒”依次走到“這幾個地方”一共需要多久	
<i>How much time will it take from “here”, to “these places” in sequence?</i>	
Speech recognition hypotheses	
1	從“這兒”依次走到“這幾個地方”你一共需要多久
2	從“這兒”依次走到“這幾個地方”給共需要多久
.....	
47	從“這兒”依次走到“這幾個地方”百移共需要多久
48	鐘“這兒”依次走到“這幾個地方”你提供需要多久
.....	
68	從“這兒”依次走到最地鐵“這裡”往你一共需要多久
69	從“這兒”依次走到“這幾個地方”往給共需要多久
70	從“這兒”依次走到最近給“這裡”往你一共需要多久
71	鐘“這兒”依次走到“這幾個地方”你移共需要多久
72	從“這兒”依次走到給地鐵“這裡”往你一共需要多久
.....	
77	從“這兒”依次走到最地鐵“這裡”往給共需要多久
78	從“這兒”依次走到“這幾個地方”米及共需要多久
79	從“這兒”依次走到“這幾個地方”裡及共需要多久
80	從“這兒”依次走到最近給“這裡”往給共需要多久
81	從“這兒”依次走到給地鐵“這裡”往給共需要多久
.....	

98	從“這兒”依次走到“這幾個地方”問一共需要多久
99	從“這兒”依次走到完“這幾個地方”給公交多久
100	從“這兒”依次走到“這幾個地方”細移公交多久

Remarks In this example, the first SLR has been transcribed as “這兒” (i.e. *here*) for 100 times across  $N$ -best speech recognition hypotheses ( $N = 100$ ). Therefore, its cost is

$$C_S(S_r[N]) = 1 - \frac{n_{S_r[N]}}{N} = 1 - \frac{100}{100} = 0$$

The second SLR has been transcribed as “這幾個地方” (i.e. *these places*) or “這裡” (i.e. *here*) for 94 and 6 times respectively across  $N$ -best speech recognition hypotheses. Therefore, the cost for “這幾個地方” (i.e. *these places*) is

$$C_S(S_r[N]) = 1 - \frac{n_{S_r[N]}}{N} = 1 - \frac{94}{100} = 0.06$$

and the cost for “這裡” (i.e. *here*) is

$$C_S(S_r[N]) = 1 - \frac{n_{S_r[N]}}{N} = 1 - \frac{6}{100} = 0.94$$

Table 6.1 An example showing the normalized cost of each recognized SLR based on Equation 6.1 for the  $N$ -best ( $N = 100$ ) recognition hypotheses

## 6.2 Filtering and Scoring the Recognized Pen Inputs

We find that subjects tend to repeat a pen gesture in referring to a location until it is highlighted on screen. We have designed a filtering mechanism to remove the repetitions. The filtering mechanism references the time and distance between two gestures as follows

**point** If a pen gesture shows the  $x$  and  $y$  coordinates within a short amount of time and a short distance, the later one is filtered out

**circle** If a pen gesture shows the four corners (i.e. maximum and minimum values of  $x$  and maximum and minimum values of  $y$ ) within a short amount of time and a short distance (as illustrated in Figure 6.2), the

later one is filtered out.

**stroke** If both of the endpoints (i.e. one for pen up and one for pen down) show the  $x$  and  $y$  coordinates within a short amount of time and a short distance, the later one is filtered out.

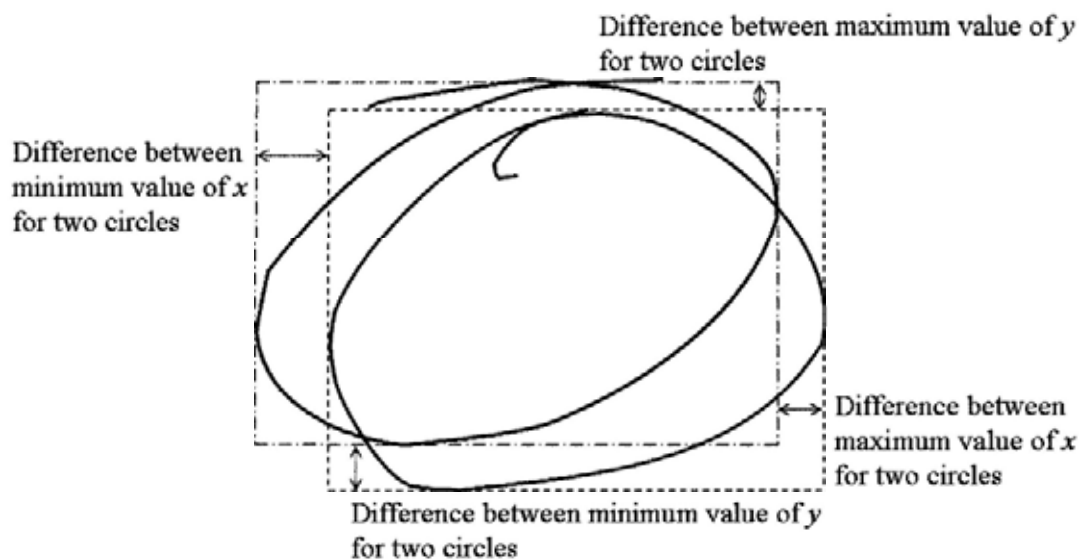


Figure 6.2: An illustration on the comparison between the four corners (i.e. maximum and minimum values of  $x$  and maximum and minimum values of  $y$ ) of two circles.

The simple pen gesture recognition algorithm mentioned in Section 5.3.2 can only generate a single output hypothesis. In order to generate the  $M$ -best pen gesture hypotheses,<sup>2</sup> we relax the constraints in pen gesture recognition and generate all possible pen gesture types as the pen gesture hypotheses.

A multimodal input expression may be transcribed as a sequence of  $Q$  pen gestures with recognized pen gesture type. Each is interpreted as a list of hypothesized locations, i.e.  $P_q$  for the  $q^{\text{th}}$  pen gesture in the input expression. The interpretations are based on locations on the map that lie within

<sup>2</sup>The maximum value of  $M$  (i.e. the maximum number pen gesture hypotheses generated) is 4 in this work.



a maximum distance  $d_{\max}$  (empirically set at 42 pixels based on the training data) from the coordinates of the pen gesture and are rank ordered based on these distances  $d_{k,q,m}$ , where  $k$  indexes the hypothesized locations in  $P_q$  and may range from 1 to  $K_{q,m}$ ; and  $m$  indexes the hypothesized pen gesture types generated by the pen gesture recognizer and  $M = 4^3$  in current work. To score a particular interpretation  $P_q[m, j]$  in the hypothesized list of the  $m$  hypothesized pen gesture type of interpreted pen gesture  $q$ , we define the normalized cost of interpretation for the pen modality  $C_P(P_q[m, j])$  as shown in Equation 6.2. The smaller the distance  $d_{j,q,m}$ , the lower the normalized cost  $C_P(P_q[m, j])$  and the more preferable the interpretation for the pen gesture. The normalized costs of the  $K_{q,m}$  hypothesized locations in  $P_q$  will sum to 1. Table 6.2 shows an illustrative example of the normalized costs of different interpretations of a pointing gesture. Hypothesized locations for the circle must have their coordinates enclosed by the circle. The locations are rank ordered based on their distances away from the circle's center.

$$C_P(P_q[m, j]) = \frac{d_{j,q,m}}{K_{q,m} \sum_{k=1}^{K_{q,m}} d_{k,q,m}} \quad (6.2)$$

where  $d_{k,q,m}$  is the distance between the coordinates of the pen gestures and hypothesized location  $k$

$K_{q,m}$  is the total number of hypothesized locations for pen gesture  $q$  within a maximum distance of  $d_{\max}$

$M$  is the total number of pen gesture type hypotheses recognized by our pen gesture recognizer

---

<sup>3</sup>We choose  $M = 4$  because the pen gesture recognizer can generate four pen gesture type, including POINT, CIRCLE, STROKE and MULTI-STROKE.

<b>Reference:</b>	
S “這兒” 的 開放時間	<i>Opening hours of “here”</i>
P. •	
<b>List of hypothesized locations:</b>	$d_{k,q,m}$ (in pixels)
1 雙秀公園 <i>Shuangxiu Park</i>	$d_{1,1,1} = 0$
2 新街口外大街 <i>Xinjiekou Outer Street</i>	$d_{2,1,1} = 9$
3 北三環 <i>North Third Ring Road</i>	$d_{3,1,1} = 10$
4 北太平橋 <i>North Taiping Bridge</i>	$d_{4,1,1} = 11$
5 北太平莊路 <i>North Taipingzhuang Road</i>	$d_{5,1,1} = 28$
6 北京師範大學 <i>Beijing Normal University</i>	$d_{6,1,1} = 46$
$\sum_{k=1}^{K_{q,m}} d_{k,q,m} = \sum_{k=1}^{K_{q,1}=6} d_{k,1,1} = 0+9+10+11+28+46 = 104$	
$C_P(P_q[1, 1]) = \frac{0}{104} = 0$	
$C_P(P_q[1, 2]) = \frac{9}{104} = 0.09$	
$C_P(P_q[1, 3]) = \frac{10}{104} = 0.1$	
$C_P(P_q[1, 4]) = \frac{11}{104} = 0.1$	
$C_P(P_q[1, 5]) = \frac{28}{104} = 0.27$	
$C_P(P_q[1, 6]) = \frac{46}{104} = 0.45$	

Table 6.2: An illustrative example for the calculation of normalized cost for the top-scoring (i.e.  $m = 1$ ) recognized pen gesture. In this example, the multimodal is transcribed as a sequence of one pen gesture (i.e.  $Q = 1$ ) and there are six hypothesized locations in total (i.e.  $K_{q,1} = 6$ )

### 6.3 Pruning and Scoring Cross-Modality Integrations

The cross-modality integration procedure described in Chapter 5 incorporates a simple cost function for the Viterbi algorithm that penalizes for mismatches in directly referenced locations, `LOC_TYPE` and `NUM` features. High accuracies in cross-modality alignment were obtained based on near-perfect multi-modal input transcriptions. However, in handling the imperfect  $N$ -best speech recognition and  $M$ -best pen recognition outputs, we need to enforce tighter constraints on semantic compatibility. We have established via the perplexity measure (in Section 5.2) that direct references should be *semantically redundant* with the corresponding pen gestures. Additionally, indirect references should be *semantically compatible* with their corresponding pen gestures.

Hence we propose to incorporate a *pruning mechanism* for candidate integrations which involve mismatches in locations between interpreted pen gestures and direct references in speech, or mismatches in the `LOC_TYPE` and `NUM` features between interpreted pen gestures and indirect references in speech. Table 6.3 presents an illustrative example. The top-scoring speech recognition hypothesis contains the direct reference 北醫 (*BMU, Beijing Medical University*) while the second best contains 北郵 (*BUPT, Beijing University of Post and Telecommunications*) instead. However, since the corresponding pen gesture (first gesture) is a point with positional coordinates that coincide with the BUPT icon (such that the distance  $d_1 = 0$ ), cross-modality integration prunes the top-scoring speech recognition hypothesis and selects the second-ranking speech recognition hypothesis due to its compatibility with pen gesture.

Candidate integrations that survive the pruning mechanism will each have a Viterbi alignment cost  $C_A(S_R, P_Q)$  (see Chapter 5, Table 5.1).  $S_R$  is the hypothesized transcription of the speech input that contains  $R$  recognized spoken locative references.  $P_Q$  is the hypothesized transcription of the pen input that contains  $Q$  interpreted pen gestures. We define the normalized

<p>Top-scoring (<math>n = 1</math>) speech recognition hypothesis (<i>pruned because of the mismatch in location between the first interpreted pen gesture and the first direct SLR</i>)</p> <p>S: 出“北醫”要到“地大”到“北科大”到“北航”最後到哪兒“北醫”坐什麼車</p> <p>P:     •           •           •           •           •</p>	
<p>Second best (<math>n = 2</math>) speech recognition hypothesis</p> <p>S: 出“北郵”到“地大”到“北科大”到“北航”最後到哪兒“北醫”坐什麼車</p> <p>P:     •           •           •           •           •</p>	
<p>Interpretation of the first pen gesture</p> <p>Point</p> <p>北京郵電大學 <i>Beijing University of Post and Telecommunications</i>     <math>d_1 = 0</math></p> <p>西土城路 <i>West Tucheng Road</i>     <math>d_2 = 18</math></p> <p>學院南路 <i>Xueyuan South Road</i>     <math>d_3 = 27.79</math></p> <p>北京師範大學 <i>Beijing Normal University</i>     <math>d_4 = 31</math></p>	

Table 6.3: An illustrative example of the pruning mechanism for candidates for cross-modality integrations. The first SLR of the top-scoring speech recognition hypothesis is the abbreviated name of “*Beijing Medical University*” while the first SLR of the second-best speech recognition hypothesis is the abbreviated name of “*Beijing University of Post and Telecommunications.*”

cost of integration  $C_I(S_R, P_Q)$ , where the subscript  $I$  denotes “integration”, as shown in Equation 6.3.  $\max\{C_A\}$  is the maximum possible Viterbi alignment cost that is empirically obtained from training data.

$$C_I(S_R, P_Q) = \frac{C_A(S_R, P_Q)}{\max\{C_A\}} \quad (6.3)$$

## 6.4 Rescoring Cross-Modality Integrations

Recall that in the current work, the speech recognizer is set to generate  $N$ -best hypotheses ( $N = 100$ ) and the pen gesture recognizer generates  $M$ -best pen gesture hypotheses ( $M = 4$ ). Cross-modality integration begins with a pruning process:

For each candidate, we apply cross-modality integration to its pair of hypothesis lists  $(S_R, P_Q)$ . Should these include incompatible semantics, the candidate is pruned (see Section 6.3 for details).

Surviving candidates (i.e. pairs of recognized speech and recognized pen hypothesis) are rescored with the following procedures:

1. If the candidate survives, we compute its normalized cost of integration  $C_I(S_R, P_Q)$  based on Equation 6.3.
2. We focus on the hypothesized transcription of the pen input  $P_Q$ . For each of the  $Q$  interpreted pen gestures (indexed by  $q$ ), we select the interpretation  $j_q$  that is semantically compatible with its aligned SLR and compute the normalized cost of pen interpretation  $C_P(P_q[m, j_q])$  (see Equation 6.2). Should there be multiple semantically compatible interpretations, their normalized costs are summed. The overall cost of interpreted pen gestures for  $P_Q$  is defined as:

$$C_P(P_Q) = \frac{1}{Q} \sum_{q=1}^Q C_P(P_q[m, j]) \quad (6.4)$$

3. We focus on the hypothesized transcription of the speech input  $S_R$ . For each of the  $R$  recognized SLRs (indexed by  $r$ ), we compute its normalized cost of recognized SLR, i.e.  $C_S(S_r[N])$  (see Equation 6.1), which is derived from the  $N$ -best purity. The overall cost of recognized SLR for SR is defined as:

$$C_S(S_R) = \frac{1}{R} \sum_{r=1}^R C_S(S_r[N]) \quad (6.5)$$

4. The rescoring function that is used to evaluate each candidate for cross-modality integration is a linear combination of the three normalized cost functions relating to the alignment, interpreted pen gestures and recognized SLRs, i.e.

$$C_{Tot}(S_R, P_Q) = w_I C_I(S_R, P_Q) + w_P C_P(C_P) + w_S C_S(S_R) \quad (6.6)$$

where  $0 < w_I, w_P, w_S < 1$  and  $w_I + w_P + w_S = 1$

We select values for the weights  $w_I$ ,  $w_P$  and  $w_S$ , by grid search to maximize cross-modality alignment accuracies based on the training data. The values selected are  $w_I = 0.5$ ,  $w_P = 0.35$  and  $w_S = 0.15$ . The “optimized” weight of the pen modality is higher than that of speech modality, possibly due to higher pen gesture recognition accuracies, as compared with the speech recognition accuracies. All candidates for cross-modality integration are rescored according to Equation 6.6 and re-ranked in ascending order of scores. As mentioned in Section 6.1, a recognition transcript with a small value of perplexity is more likely to have a reasonable interpretation. Therefore, if there is a tie in the scores after re-ranking, the candidates will be ranked in ascending order of their perplexity. The candidate with minimum overall cost  $C_{Tot}(S_R, P_Q)$  is identified as the preferred cross-modality alignment. An illustrative example is shown in Table 6.4 and complete list of speech recognition hypotheses and their overall costs  $C_{Tot}(S_R, P_Q)$  is shown in Table E.1 of Appendix E.

<b>Reference transcription</b>
S: 從“這兒”依次走到“這幾個地方”一共需要多久
P:     •                                     •••••
<i>How much time will it take from here, to these places in sequence?</i>

SR rank	Hypothesis Pairs and $C_{Tot}(S_R, P_Q)$	HR rank
1	<p>S: 從“這兒”依次走到“這幾個地方”你一共需要多久</p> <p>P:     •                                     •••••</p> $C_{Tot}(S_R, P_Q) = w_I C_I(S_R, P_Q) + w_P C_P(P_Q) + w_S C_S(S_R)$ $= 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 23.03</math></p>	8
2	<p>S: 從“這兒”依次走到“這幾個地方”給共需要多久</p> <p>P:     •                                     •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 30.89</math></p>	20
.....		
47	<p>S: 從“這兒”依次走到“這幾個地方”百移共需要多久</p> <p>P:     •                                     •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 45.36</math></p>	36
48	<p>S: 鐘“這兒”依次走到“這幾個地方”你提供需要多久</p> <p>P:     •                                     •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 124.07</math></p>	88
.....		

68	<p>S: 從 “這兒” 依次 走到 最 地鐵 “這裡” 往 你 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0 \ 94)}{2} = 0.0705$ <p><math>PP_{MM} = 79.69</math></p>	95
69	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 往 給 共需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0 \ 06)}{2} = 0.0045$ <p><math>PP_{MM} = 51.11</math></p>	45
70	<p>S: 從 “這兒” 依次 走到 最近 給 “這裡” 往 你 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0 \ 94)}{2} = 0.0705$ <p><math>PP_{MM} = 102.15</math></p>	98
71	<p>S: 鐘 “這兒” 依次 走到 “這幾個地方” 你 移 共需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0 \ 06)}{2} = 0.0045$ <p><math>PP_{MM} = 115.51</math></p>	86
72	<p>S: 從 “這兒” 依次 走到 給 地鐵 “這裡” 往 你 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0 \ 94)}{2} = 0.0705$ <p><math>PP_{MM} = 84.15</math></p>	96
.....		
77	<p>S: 從 “這兒” 依次 走到 最 地鐵 “這裡” 往 給 共需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0 \ 94)}{2} = 0.0705$ <p><math>PP_{MM} = 99.8442</math></p>	97



78	<p>S: 從“這兒”依次走到“這幾個地方”米及共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 42.46$	29
.....		
79	<p>S: 從“這兒”依次走到“這幾個地方”裡及共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 55.80$	48
80	<p>S: 從“這兒”依次走到最近給“這裡”往給共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 127.99$	100
81	<p>S: 從“這兒”依次走到給地鐵“這裡”往給共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 105.44$	99
.....		
98	<p>S: 從“這兒”依次走到“這幾個地方”問一共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 3.03$	1
99	<p>S: 從“這兒”依次走到完“這幾個地方”給公交多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 81.89$	71

100	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 細 移 公交 多久</p> <p>P:           •                                  • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 115.16$	85
-----	--	----

Table 6.4: An example illustrating the hypothesis rescoring process of based on the  $N$ -best speech recognition hypotheses ( $N = 100$ ) listed in Table 6.1. The second SLR, *these places*, should have NUM=plural, which can be aligned with more than one pen gestures. Another possibility of the second SLR is *here*, which should have NUM=nil and can be aligned with any number of pen gestures. All the five pen gestures incur no cost because their coordinates coincide with the respect icons/labels. Each candidate for cross-modality integration is rescored and then the updated rank is shown for each candidate. The 98<sup>th</sup> hypothesis pair ranked top after rescoring.

### 6.5 Evaluating the Rescoring Procedure

The application of the rescoring procedure to the candidate hypotheses for cross-modality integration has brought some improvements to the alignment accuracies in the training and test sets of our multimodal corpus. Table 6.5 and Figure 6.3 summarizes the results of the percentage of correctly aligned expressions. These are expressions for which our framework can generate unimodal verbalized paraphrases that are semantically equivalent with the original multimodal expressions. Improvement in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from 51.1% to 71.8% for the training set and from 54.4% to 72.8% for the test set. Further analysis of our results (see Table 6.6) shows that there can be correct cross-modality integration despite recognition errors in speech and/or pen modalities. The  $N$ -best hypothesis rescoring framework can effectively re-rank the hypothesis pairs to obtain correct integration, as illustrated by the

examples in Table 6.7

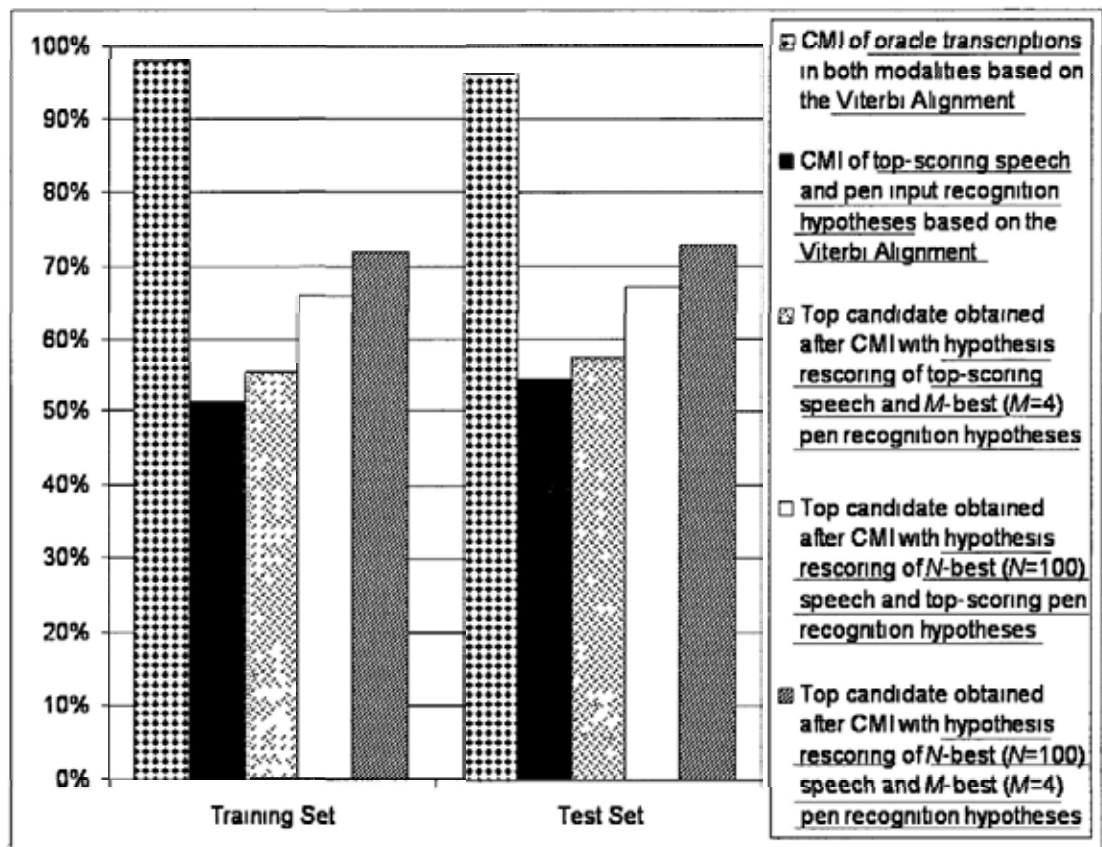


Figure 6.3 Performance of cross-modality integration (CMI) in the training and test sets

In addition, analysis of the incorrect alignments (as shown in Table 6.8) after re-scoring and re-ranking suggests that the incorporation of finer cross-modality timing information will be helpful. Such timing information should be used judiciously since the modalities are not necessary simultaneous and user's integration pattern may vary during the interaction [51]. A possible extension on this work is to detect the subject's integration pattern (i.e. simultaneous integrator or sequential integrator) and incorporate the timing information in the semantic integration framework for simultaneous integrator.

Furthermore, a good number of the errors are associated with the recognized SLR 這兒 (i.e. *here*) having an unspecified NUM feature and can thus be aligned with an arbitrary number of pen gestures. Making the assumption of NUM= 1 should be helpful for error recovery. This is because analysis of the cross-modal integration patterns in the *manually transcribed* training set (see Table 4.8) shows that 94% of the unspecified reference (i.e. *here*) is used to referring single location. Application of this assumption (i.e. 這裡 or *here* has NUM=1) shows that although it cannot improve the performance of the cross-modality semantic integration framework, it can provide a more specific alignment cost for the hypothesis pairs for rescoring as shown in Table 6.9.

## 6.6 Chapter Summary

In this chapter, we present a hypothesis rescoring framework pertaining to achieve robustness towards imperfect transcripts. For each multimodal expression, this procedure considers all candidates for cross-modality integration based on the  $N$ -best ( $N = 100$ ) speech recognition hypotheses and the  $M$ -best ( $M = 4$ ) pen input recognition hypotheses. Note that the single recognized pen gesture can generate  $Q$  location hypotheses that are fed into the cross-modality hypothesis rescoring procedure (see Equation 6.2). Rescoring combines such elements as the integration scores obtained from the Viterbi algorithm,  $N$ -best purity for recognized spoken locative references, as well as distances between coordinates of recognized pen gestures and relevant icons on the map. Experiments using the  $N$ -best ( $N = 100$ ) speech recognition hypothesis and top-scoring ( $M = 4$ ) pen recognition hypotheses show that the rescoring and re-ranking helped improve the performance of correct cross-modality interpretation from 51.1% to 71.8% for the training set and from 54.4% to 72.8% for the test set.

	Training Set	Test Set
Number of multimodal inquiries	1002	440
Number of multimodal inquiries that contain SLR(s)	968	434
Cross-modality integration of oracle transcriptions in both modalities based on the Viterbi alignment in Chapter 5	98.1% (950/968)	95.9% (416/434)
Cross-modality integration of top-scoring speech and pen gesture recognition hypotheses based on the Viterbi alignment in Chapter 5	51.1% (495/968)	54.4% (236/434)
Top candidate obtained after cross-modality integration and rescoring of the top-scoring speech recognition hypothesis and $M$ -best pen recognition hypotheses ( $M = 4$ )	55.3% (535/968)	57.6% (250/434)
Top candidate obtained after cross-modality integration and rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) with the top-scoring pen gesture recognition hypothesis	65.9% (638/968)	67.1% (291/434)
Top candidate obtained after cross-modality integration and rescoring of the $N$ -best speech recognition hypotheses ( $N = 100$ ) with the $M$ -best pen recognition hypotheses ( $M = 4$ )	71.8% (695/968)	72.8% (316/434)

Table 6.5: Performance of cross-modality integration, measured in terms of the percentage of correctly aligned expressions in the training and test sets. Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from (1) top-scoring hypotheses from speech and pen to top-scoring hypotheses from speech and  $M$ -best hypotheses from pen; (2) top-scoring hypotheses from speech and pen to  $N$ -best hypotheses from speech and top-scoring hypotheses from pen; (3) top-scoring hypotheses from speech and pen to  $N$ -best hypotheses from speech and  $M$ -best hypotheses from pen ( $\alpha = 0.01$ , two-tailed  $z$ -test) as shown in Appendix F.

Pen recog- nition	SLR recog- nition	Number of in- quiries in the test set (440 in total)	Correct integra- tion with top- scoring hypothe- ses from each modality	Correct integra- tion with $N$ -best ( $N = 100$ ) speech recognition hypo- theses and $M$ -best ( $M = 4$ ) pen recognition hypotheses
Correct	Correct	98/434 (22.6%)	98/98 (100%)	98/987 (100%)
Correct	Incorrect	260/434 (59.9%)	98/260 (37.7%)	159/260 (61.1%)
Incorrect	Correct	42/434 (9.6%)	29/42 (68.3%)	39/42 (92.7%)
Incorrect	Incorrect	34/434 (7.9%)	11/34 (32.4%)	20/34 (58.8%)
Overall			54.4%	72.8%

Table 6.6: Detailed performance statistics of the test set. Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant in the presence of speech and/or pen recognition errors ( $\alpha = 0.01$ , two-tailed  $z$ -test) as shown in Appendices F.4, F.5 and F.6.

**Example 1 (with SLR recognition errors):**

Reference transcriptions:

S: 從“這兒”到“這四個大學”要多久？

P:       •           ••••

*How much time will it take from “here” to “these four unversities?”*

Top-scoring speech and pen recognition hypotheses:

S: 從“這裡”到“這些地方”要多久？

P:       •           ••••

*how much time will it take from “here” to “these locations”?*

*Remark: the reference SLR, 這裡, has the same semantic meaning as 這兒 (i.e. “here”) and does not affect the subsequent cross modality integration. The numeric and the location type features are lost during recognition of the second SLR “these locations”. The proposed framework can find out the correct alignment and extract the name of the four universities based on the complementary relation between the modalities.*

**Example 2 (with pen gesture recognition error):**

Reference transcriptions:

S: 在“這裡”逛一圈要多久？

P:       ○ (user drew a flat circle to indicate a street)

*how long will it take to stroll around “here”?*

Top-scoring speech and pen recognition hypotheses:

S 在“這裡”逛一圈要多久？

P:       → (pen gesture mis-recognized as a stroke to indicate a street)

*how long will it take to stroll around “here”?*

*Remark: The pen interpretation method in Section 4.4 can identify the street as indicated by the mis-recognized pen gesture and hence the recognition error does not affect the cross-modality semantic integration process.*

<p><b>Example 3 (with SLR and pen gesture recognition errors):</b></p> <p>Reference transcriptions</p> <p>S: “這個公園” 什麼 時候 開放</p> <p>P: • (a big point within the icon of a park)</p> <p><i>what is the opening hours of “this park”?</i></p>
<p>Top-scoring speech and pen recognition hypotheses:</p> <p>S: “這兒” 公園 什麼 時候 開放</p> <p>P: ○ (a circle within the icon of a park)</p> <p><i>what is the opening hours of “here” park?</i></p>
<p><i>Remark: Although the numeric and location type features are missed in the recognized SLR and the point is mistaken as circle by the pen gesture recognizer, the framework can integrate the two modalities correctly and identify the park indicated by the user.</i></p>

Table 6.7: Examples on the correct integration with the present of SLR and/or pen recognition error.



<p><b>Reference transcription:</b></p> <p>S. 我在“這裡”從“這裡”要到“這些地方”一共需要多久</p> <p>P.           •                           •••••</p> <p><i>I'm now at "here". How much time will it take from "here" to "theses places"?</i></p>
<p><b>Result of Cross-Modality Integration after Hypothesis Rescoring:</b></p> <p>S. 我在“這裡”從“這裡”要到“這些地方”一共需要多久</p> <p>P.           •           •                   •••</p>

Table 6.8: An example on the incorrect alignment due to the presence of NUM=plural (from “these places”) and missing of timing information during integration. Since NUM feature of *these places* is plural, which can align with *more than one pen gestures* without a specific number, our framework align one of the pen gesture to the anaphora (i.e. the second “here”) in the spoken input.

<p><b>Reference transcription:</b></p> <p>S. 到“這些地方”和“這裡”有什麼路線選擇</p> <p>P:           •••           •</p> <p><i>Visit "these places" and "here". What are the possible routes?</i></p>
<p><b>Result of Cross-Modality Integration after Hypothesis Rescoring:</b></p> <p>S. 到“這些地方”和“這裡”有什麼路線選擇</p> <p>P:           ••                   ••</p>

Table 6.9: An example on the incorrect alignment due to the presence of unspecified NUM feature (i.e. NUM=nil). Since NUM feature of *here* is unspecified, it can align with *any arbitrary number of pen gestures* without penalty.

## Chapter 7

# Latent Semantic Analysis for Multimodal User Input with Speech and Pen Gestures

This chapter describes our attempt in developing a semantic analysis framework for multimodal user input with speech and pen gestures. More specifically, our aim is to infer the domain-specific task goal(s) of the multimodal input. The task goal is characterized by terms used in the spoken modality, as well as particular term co-occurrence patterns across modalities. Previously, we have applied Belief Networks [71] [72] for task goal inference based on unimodal (speech-only) inputs. Since multimodal input usually has a simpler syntactic structure than unimodal input [73] and the order of semantic constituents is different in multimodal and unimodal input [17], we apply latent semantic modeling (LSM) in capturing the latent semantics of the multimodal user inputs as well as the task goals. As such, LSM is a data-driven approach that models the underlying semantics of word usages from available corpora. It has been applied unimodally to text or transcribed speech for language modeling [74], document clustering [75], spoken document retrieval [76], document

summarization [77], etc. This is carried out with the objective of uncovering the associations between (unimodal or multimodal) terms and task goals through a data-derived latent space.

In LSM, the association between terms (including both lexical and multimodal terms) and task goals is represented as a term-task goal matrix. This can be factorized into a term-semantics and a task goal-semantics matrix using singular value decomposition (SVD). These two matrices associate terms and task goals through an automatically derived space of semantics, instead of directly relating the terms with task goals. Based on the latent semantic space, we can reconstruct the space of terms and task goals. We can then examine the structural relations between terms and task goals in the reconstructed space. There are a total of nine task goals in our experimental domain. In the following, we introduce latent semantic analysis, present the collected multimodal corpus and discuss the process of task goal inference and related experimentation.

## 7.1 Latent Semantic Modeling of Cross-modal Integration Patterns

We apply latent semantic modeling (LSM) [78] to capture regularities in terms (including both lexical and multimodal terms) from a multimodal expression, in relation to their usage contexts (i.e. task goal in this work). LSM uses singular value decomposition (SVD) to derive a latent semantic space that relates terms (combined lexical, gestural and multimodal terms<sup>17</sup>) with the task goals.

---

<sup>17</sup>Examples of lexical terms are 多久 (i.e. *how long*), 公車 (i.e. *bus*), examples of gestural terms are  $\langle \emptyset \mid \text{POINT} \mid \emptyset \rangle$  and  $\langle \emptyset \mid \text{CIRCLE} \mid \emptyset \rangle$ , and examples of multimodal terms are  $\langle \text{這條街} \mid \text{STROKE} \mid \text{SIM} \rangle$  (i.e.  $\langle \text{this street} \mid \text{STROKE} \mid \text{SIM} \rangle$ ) and  $\langle \text{這個範圍} \mid \text{CIRCLE} \mid \text{SEQ} \rangle$  (i.e.  $\langle \text{this area} \mid \text{CIRCLE} \mid \text{SEQ} \rangle$ )

### 7.1.1 Association Matrices

Associations between terms (including both lexical and multimodal terms) and task goals can be summarized in a term-task goal matrix  $B$ . Given  $M$  terms (details of the multimodal terms are presented in Section 4.6) and  $A$  task goals, we form  $M \times A$  matrix  $B$ . Each row represents a term. Each column represents a task goal. The element  $b_{m,a}$ , is the weight (i.e. normalized term frequency using term frequency-inverse document frequency (TF-IDF)<sup>18</sup>) [79] for the term  $m$  in the  $a^{\text{th}}$  task goal. The training set consists of 881 terms (i.e.  $M = 881$ ). The statistics of lexical and multimodal terms in the training set are shown in Table 7.1.1). There are nine task goals (i.e.  $A = 9$ ) in this work.

$$B = \begin{bmatrix} b_{1,1} & \dots & b_{1,a} & \dots & b_{1,A} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{m,1} & \dots & b_{m,a} & \dots & b_{m,A} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{M,1} & \dots & b_{M,a} & \dots & b_{M,A} \end{bmatrix} \quad (7.1)$$

---

<sup>18</sup>The term frequency (TF) can be used to indicate the term importance with the assumption that frequent terms are more important. The inverse document frequency (IDF) can be used to discount non-discriminative terms, e.g. function words of 的 (i.e. *of*), 啊 (i.e. *oh*) and 是 (i.e. *is*), etc. This is based on an assumption that discriminating power of a term decreases with the number of times that the term occurs in the data set.

where  $b_{m,a} = (1 - \varepsilon_m) \frac{\kappa_{m,a}}{\lambda_a}$ ,

$$\varepsilon_m = -\frac{1}{\log A} \sum_{a=1}^A \frac{\kappa_{m,a}}{\tau_m} \log \frac{\kappa_{m,a}}{\tau_m},$$

$\frac{\kappa_{m,a}}{\lambda_a}$  is the term frequency of term  $m$ ,

$\log \frac{\kappa_{m,a}}{\tau_m}$  is the inverse document frequency of term  $m$ ,

$\kappa_{m,a}$  denotes the number of times the term  $m$  occurs in the  $a^{\text{th}}$  task goal,

$\lambda_a$  is the total number of terms in the  $a^{\text{th}}$  task goal,

$\varepsilon_m$  denotes the normalized entropy of term  $m$  in the data set; and

$\tau_m$  is the total number of times that term  $m$  occurs in the training set.

(7.2)

Number of multimodal terms	567
(SLR and pen)	508
(SLR only)	53
(Pen only)	6
Number of lexical terms	314
Total number of terms	881

Table 7.1: Statistics of lexical and multimodal terms in the training set.

$B$  can be decomposed into a product of three matrices, with methods such as singular value decomposition (SVD) [78], probabilistic latent semantic analysis (PLSA) [80] and latent Dirichlet allocation (LDA) [81]. We propose to focus on the use of SVD of order  $R$ , as shown in Equation 7.3.

$$\begin{aligned}
 B &= USV^T \\
 &= \begin{bmatrix} u_{1,1} & \cdots & u_{1,R} \\ \vdots & \ddots & \vdots \\ u_{M,1} & \cdots & u_{M,R} \end{bmatrix} \begin{bmatrix} s_{1,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_{R,R} \end{bmatrix} \begin{bmatrix} v_{1,1} & \cdots & v_{1,R} \\ \vdots & \ddots & \vdots \\ v_{A,1} & \cdots & v_{A,R} \end{bmatrix}^T \quad (7.3)
 \end{aligned}$$

where  $U$  is the term-semantics matrix of dimensions  $M \times R$ ,

$S$  is the diagonal matrix of singular values sorted in descending order with dimensions  $R \times R$ ,

$V$  is task goal-semantics matrix of dimensions  $A \times R$ ,

$R = \min\{M, A\}$  is the order of decomposition and

$T$  is the transpose of the matrix.

$U$  and  $V$  are the left unitary matrix and right unitary matrix respectively. Each column of  $U$  contains the estimated weight of each term  $m$  that corresponds to the latent semantic category  $r$  while each column of  $V^T$  contains the estimated weight of each task goal  $a$  that corresponds to the latent semantic category  $r$ . Equation 7.3 projects the space of terms and task goals onto a reduced  $R$ -dimensional space which is defined by the orthonormal basis given by the column vectors  $u_m$  and  $v_a$  from matrices  $U$  and  $V$  respectively. In order to collapse the terms that are “semantically similar”, we always choose  $R' \leq R$ . The smaller the value  $R'$ , the more pronounced the reduction of semantic redundancy in the latent semantic space. Based on the latent semantic space, we may re-construct the space of terms and task goals, denoted as  $\hat{B}$  in Equation 7.4.

$$B \approx \hat{B} = U\hat{S}V^T \quad (7.4)$$

where  $\hat{S}$  is the reduced diagonal matrix of singular values with optimized value of  $R'$  (i.e. with dimensions  $R' \times R'$ ).

We need to find an “optimal” choice of  $R'$  that minimizes the distortion between the re-constructed space  $\hat{G}$  and the original space  $G$ , in the implementation of Equation 8.4 in the training procedure. Since we have nine task goals in this work and we aim to analyze the structural relations between terms and each task goal, we simply choose  $R' = R = 9$ . We re-construct the space of terms and task goals based on Equation 7.4 and examine the structural relations between terms and task goals in the reconstructed space.

## 7.2 Task Goal Inference

In Chapter 4, we examined the characteristics of SLR and pen gestures and the cross-modality associations between SLRs and pen gestures, leading to the definition of a multimodal term that captures cross-modal integration patterns and their temporal relationships. In this section, we present a framework for inferring the task goal based on an input inquiry.

### 7.2.1 Performance Baseline using Vector-Space Model

As a reference baseline, we apply the vector-space model [82] for task goal inference. For each task goal  $a$ , we consider all of its training expressions and their multimodal terms. We create a vector  $j_a$  of weights, using the normalized term frequency TF-IDF of the multimodal terms. For a task goal, we create a vector  $b_n$ , similar to the column vector of  $B$  in Equation 7.1. The similarity between  $b_n$  and  $j_a$  is calculated as the inner product of the two vectors (see Figure 7.1). Long inquiries contain more terms. Since the dot product favors long inquiries by generating higher similarity scores, we apply cosine normal-

ization (i.e. divide the dot product by the Euclidean Distance [82] [83] between the two vectors) to reduce the adverse effect of term repetition. Equation 7.5 shows the similarity calculation using the dot product between the unit vector of  $j_a$  and the unit vector of  $b_n$ .

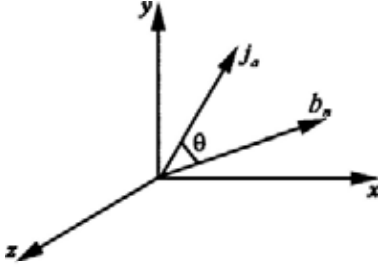


Figure 7.1: Similarity between vector  $j_a$  and  $b_n$  captured by the cosine of the angle  $\theta$  between them. The angle  $\theta$  between the two vectors is 0, corresponding to maximal similarity.

$$similarity_{cosine}(j_a, b_n) = \frac{j_a \cdot b_n}{\|j_a\| \|b_n\|} \quad (7.5)$$

where  $j_a$  is the weight for all terms in the  $a^{th}$  task goal and  $b_n$  is the weight for all terms in the  $n^{th}$  task goal.

The task goal vector is assigned to the task goal  $a_n^*$  which has the maximum similarity score, as shown in Equation 7.6.

$$a_n^* = \arg \max_a \{similarity_{cosine}(j_a, b_n)\} \quad (7.6)$$

Experiments show that the vector-space model can correctly infer task goals for all of the task goal vectors in training and test sets respectively. Table 7.2.1 shows the performance of task goal inference using vector-space model based on different weighting methods. Due to the adverse effect of non-discriminative terms, task goal inference performance based on TF (the first



two rows of Table 7.2.1) is lower than the one based on TF-IDF (the last two rows of Table 7.2.1). Moreover, due to the adverse effect of term repetition and the fact that dot product favors long inquiries, the task goal inference performance using dot product (the first and third rows of Table 7.2.1) is lower than the one using cosine normalization (the second and forth rows of Table 7.2.1). The best performance is achieved using cosine similarity based on TF-IDF.

	Training set	Test set
Dot product (without cosine normalization) based on term frequency (see Equation 7.2)	33.3% (3/9)	25% (2/8)
$similarity_{cosine}(j_a, g_n)$ (see Equation 7.5) based on term frequency	66.7% (6/9)	62.5% (5/8)
Dot product (without cosine normalization) based on TF-IDF	77.8% (7/9)	75% (6/8)
$similarity_{cosine}(j_a, g_n)$ based on TF-IDF	100% (9/9)	100% (8/8)

Table 7.2: Task goal inference accuracy using vector-space model based on different weight methods. Please note that the test set lacks expressions in task goal CHOICE OF VEHICLE (i.e. only contains 8 task goal vectors).

### 7.2.2 Performance Evaluation

Overall performance in task goal inference for the training and test sets are 100% (9/9) and 100% (8/8) respectively since the test set lacks expressions that fall under the task goal CHOICE OF VEHICLE.

### 7.2.3 Analysis of the Re-constructed Space for Identification of Key Terms

We examine the term weights in the re-constructed space to identify key terms that are indicative of each task goal. Lexical and multimodal terms with high LSM weights and the identified key terms for each task goal are shown in Table 7.2.3. Figures 7.2 to 7.10 are the plots of term weight (for both lexical and multimodal terms) from matrix  $\hat{B}$  against lexical and multimodal terms for each of the task goals. The key terms identified can be used for the understanding and interpretation of the user input.

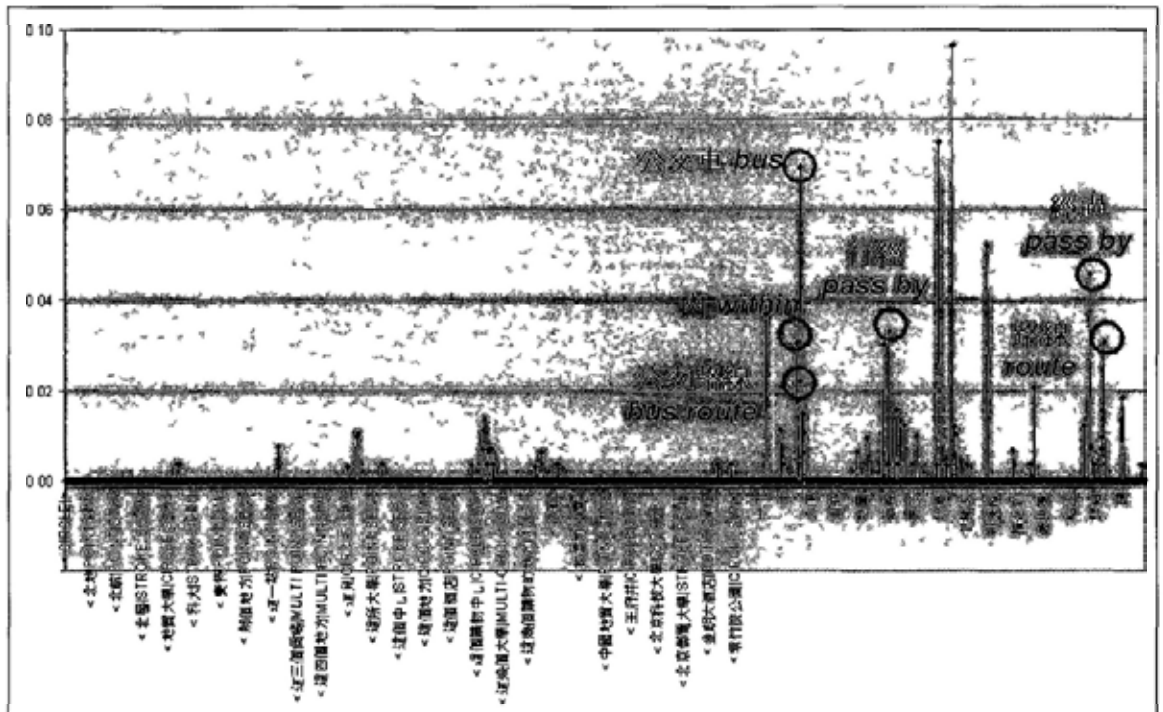


Figure 7.2: A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal BUS INFORMATION.

<p>BUS INFORMATION</p>	<p>公交車* <i>bus</i>*, 經過* <i>pass by</i>*, 行經* <i>pass by</i>*, 路線* <i>route</i>*,          公交路線* <i>bus route</i>*, 內* <i>within</i>*,          的 <i>of</i>, 所有 <i>all</i>, 哪些 <i>which</i>, 一百米 <i>100m</i>, 有 <i>have</i>, 都有 <i>have</i></p>
<p>CHOICE OF VEHICLE</p>	<p>坐* <i>take</i>*, 不是* <i>not</i>*, 公交車* <i>bus</i>*, 地鐵* <i>railway</i>*, 乘* <i>take</i>*,          公交* <i>bus</i>*, 搭* <i>take</i>*,          我 <i>I</i>, &lt;北京站   CIRCLE   SIM&gt; &lt;<i>Beijing Station</i>   CIRCLE   SIM&gt;,          希望 <i>wish</i>, 怎麼走 <i>how should I go</i>, 要在 <i>in</i>, 去 <i>go</i>, 現在 <i>now</i>,          &lt;新東安市場   POINT   SIM&gt; &lt;<i>Xindong'an Plaza</i>   POINT   SIM&gt;,          要 <i>want</i>, &lt;這裡   POINT   SEQ&gt; &lt;<i>here</i>   POINT   SEQ&gt;,          &lt;⊙   STROKE   ⊙&gt;, 在 <i>in</i></p>
<p>MAP COMMANDS</p>	<p>放大* <i>zoom in</i>*, 地圖* <i>map</i>*, 詳細* <i>detailed</i>*, 縮小* <i>zoom out</i>*,          顯示* <i>show</i>*, 畫面* <i>screen</i>*, &lt;⊙   POINT   ⊙&gt;*,          &lt;這裡   CIRCLE   SEQ&gt;* &lt;<i>here</i>   CIRCLE   SEQ&gt;*,          的 <i>of</i>, 我 <i>I</i>, 到 <i>to</i>, 請 <i>please</i>, 想 <i>wish</i>, 更 <i>still</i>, 將 <i>get</i></p>
<p>OPENING HOURS</p>	<p>幾點* <i>when</i>*, 開放時間* <i>opening hours</i>*, 時候* <i>time</i>*,          開放* <i>opening</i>*, 營運時間* <i>opening hours</i>*, 什麼* <i>what</i>*,          的 <i>of</i>, 是 <i>is</i>, 請問 <i>please</i>, 和 <i>and</i>, 我 <i>I</i>, 想 <i>wish</i></p>
<p>RAILWAY INFORMATION</p>	<p>地鐵站* <i>railway station</i>*, 多少個* <i>how many</i>*, 附近* <i>nearby</i>*,          範圍* <i>area</i>*, 名稱* <i>name</i>*, 周圍* <i>surroundings</i>*, 內* <i>within</i>*,          四百米* <i>400m</i>*, 五百米* <i>500m</i>*,          有 <i>have</i>, 的 <i>of</i>, 請問 <i>please</i>, 所有 <i>all</i>, 我 <i>I</i></p>
<p>ROUTE FINDING</p>	<p>到* <i>to</i>*, 從* <i>from</i>*, 最快* <i>the fastest</i>*, 怎麼走* <i>how should I go</i>*,          &lt;這裡   POINT   SIM&gt;* &lt;<i>here</i>   POINT   SIM&gt;*,          &lt;這兒   POINT   SIM&gt;* &lt;<i>here</i>   POINT   SIM&gt;*,          &lt;這個大學   POINT   SIM&gt;* &lt;<i>this university</i>   POINT   SIM&gt;*,          我 <i>I</i>, 的 <i>of</i></p>

TIME CONSTRAINT	二十分鐘* <i>20 mins.*</i> , 內* <i>within*</i> , 到達* <i>arrive at*</i> , 想* <i>wish*</i> , 之內* <i>within*</i> , <這裡   POINT   SIM>* <i>&lt;here   POINT   SIM&gt;*</i> , 到* <i>to*</i> , <國際飯店   POINT   SIM> <i>&lt;International Hotel   POINT   SIM&gt;</i> , 我 <i>I</i> , 在 <i>in</i>
TRANSPORTATION COSTS	多少錢* <i>how much will it cost*</i> , 到* <i>to*</i> , 地鐵* <i>railway*</i> , 從* <i>from*</i> , 坐* <i>take*</i> , 需要* <i>need*</i> , 要* <i>need*</i> , <這裡   POINT   SIM> <i>&lt;here   POINT   SIM&gt;</i> , 請問 <i>please</i>
TRAVEL TIME	到* <i>to*</i> , 從* <i>from*</i> , 多長時間* <i>how long*</i> , 要* <i>need*</i> , 需要* <i>need*</i> , 多久* <i>how long*</i> , 多少時間* <i>how long*</i> , 一共需要* <i>need in all*</i> , 再到* <i>then go*</i> , <這裡   POINT   SIM>* <i>&lt;here   POINT   SIM&gt;*</i> , <這個大學   POINT   SIM>* <i>&lt;this university   POINT   SIM&gt;*</i> , 我 <i>I</i>

Table 7.3: Lexical and multimodal terms with the highest LSM weights for each task goal. Terms with an asterisk (i.e. \*) are the identified key terms.

### 7.3 Chapter Summary

In this chapter, we have extended our study to the usage pattern and latent semantic analyzes of multimodal user inputs with speech and pen gestures. Our investigation is based on a multimodal corpus that we have designed and collected, which consists of over a thousand navigational inquiries. The inquiries cover nine task goals. The task goal of each multimodal input is hand-labeled as a gold standard. We use a non-negative term-task goal matrix to capture the associations between terms (lexical and multimodal) and task goals. Decomposition of the term-task goal matrix using singular value decomposition (SVD) captures the associations between terms and task goals through a latent semantic space. We can then reconstruct the space of terms and task

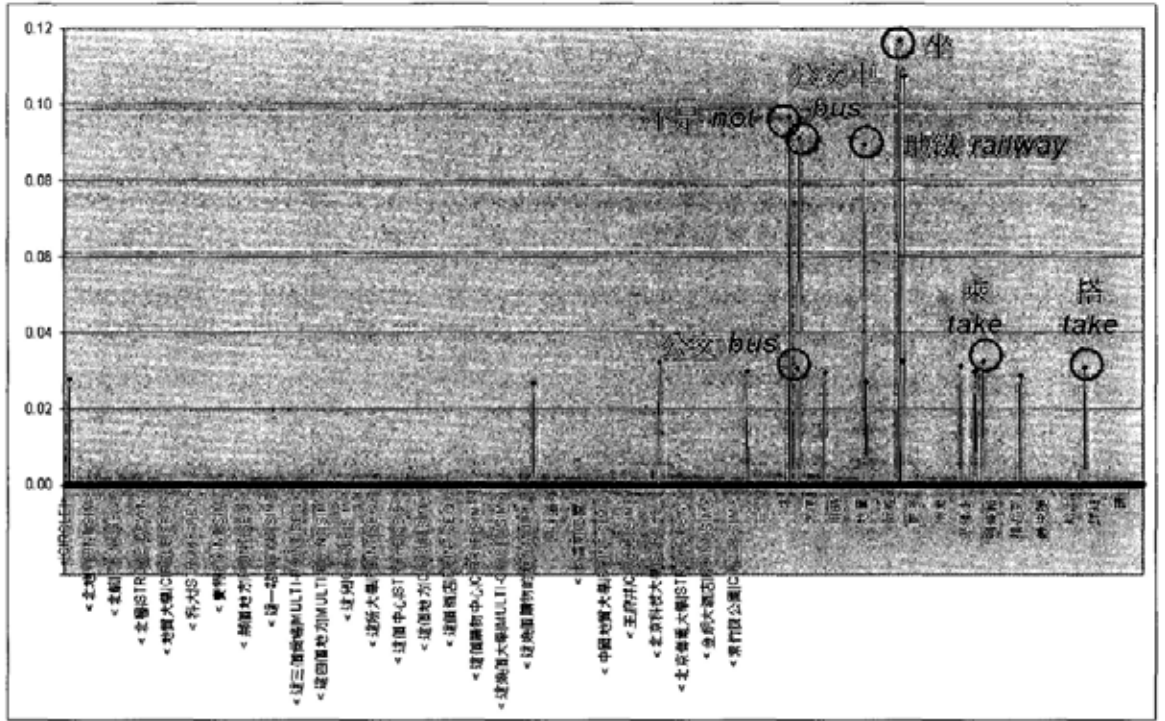


Figure 7.3: A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal CHOICE OF VEHICLE.

goals based on the latent semantic space. Examination of the term weights in the re-constructed space can identify key terms that are indicative of each task goal.

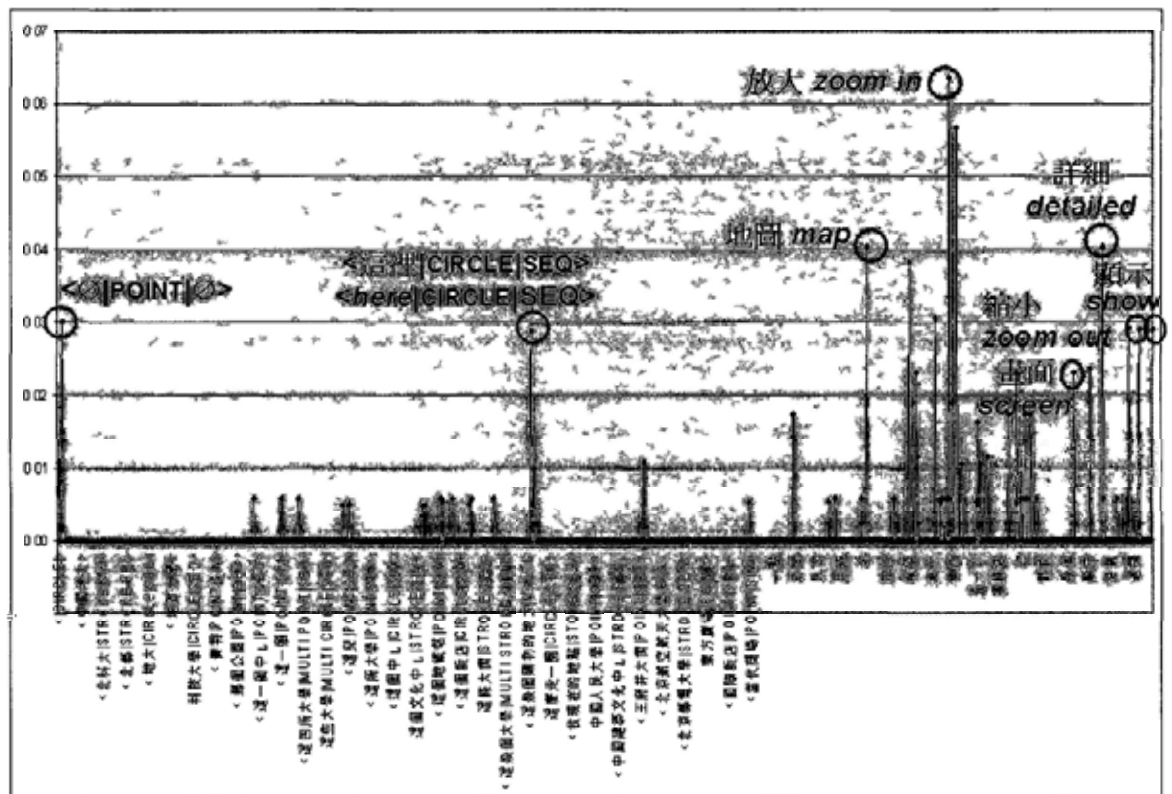


Figure 7.4 A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal MAP COMMANDS

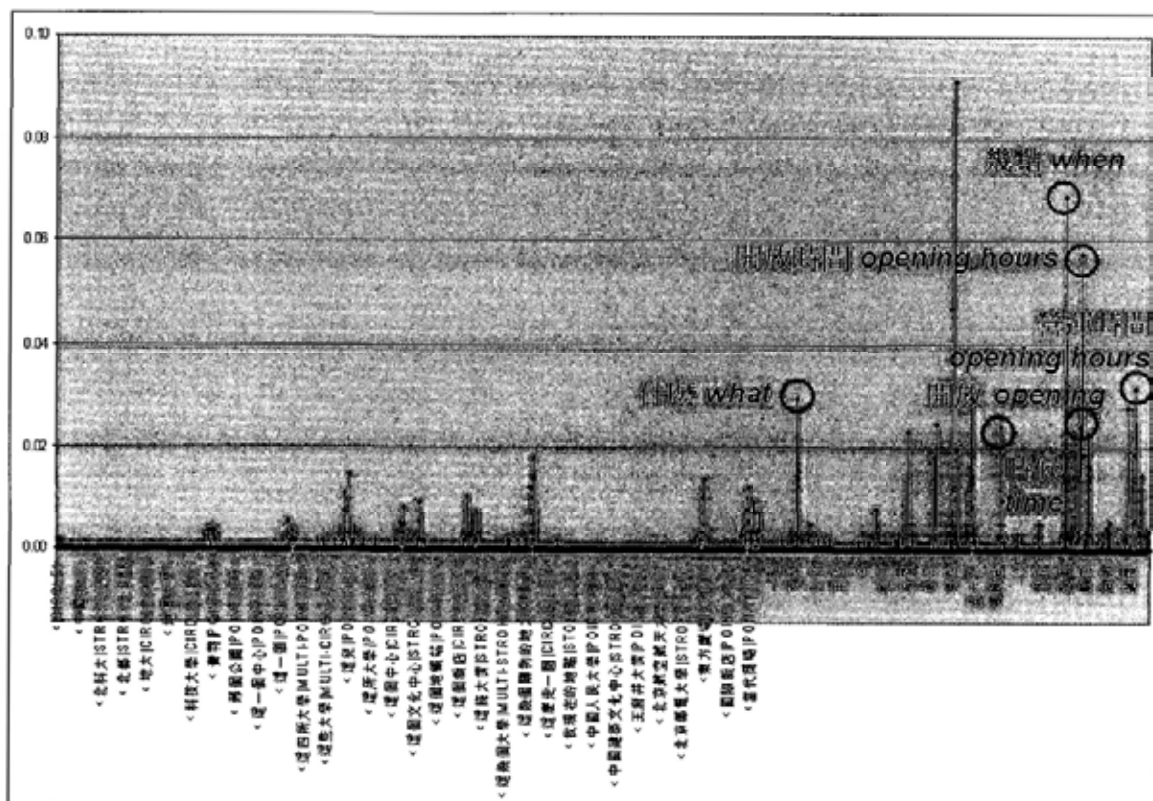


Figure 7.5: A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal OPENING HOURS.

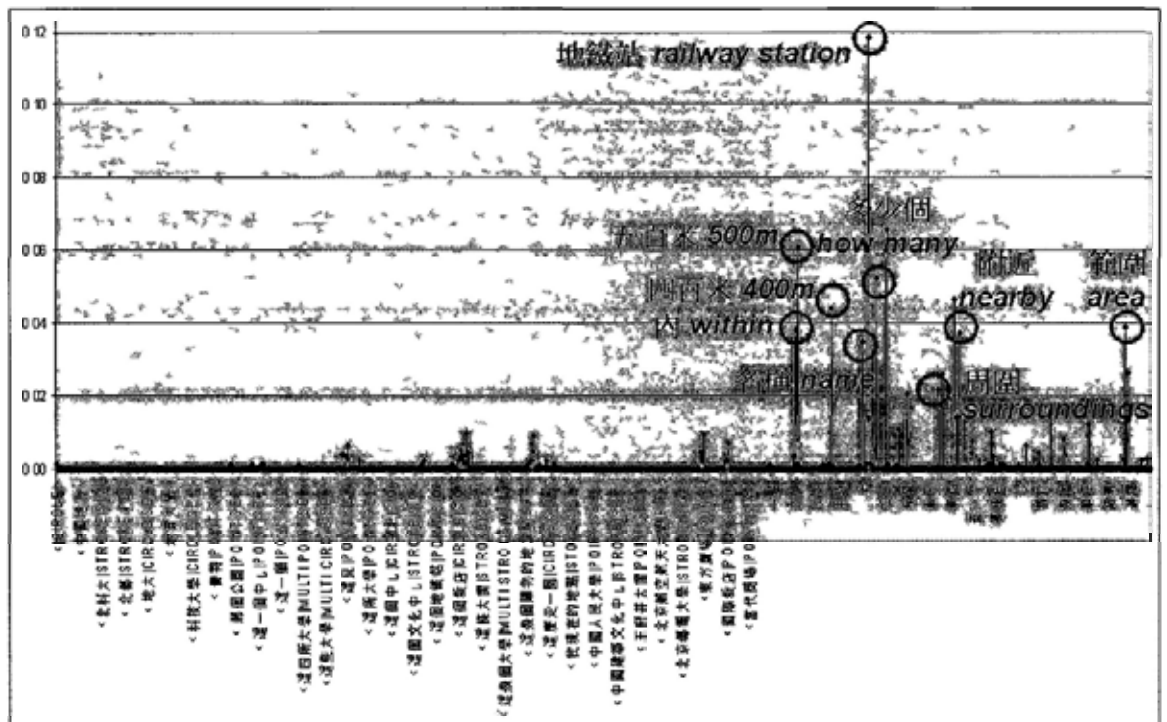


Figure 7.6 A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal RAILWAY INFORMATION



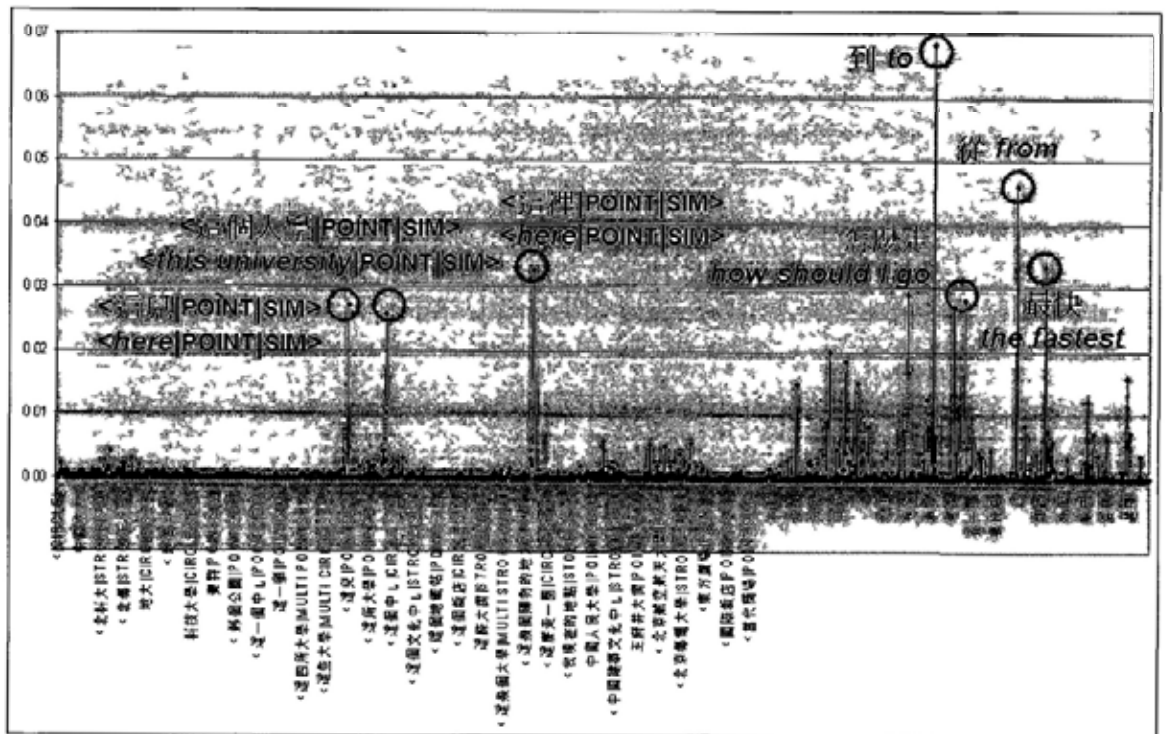


Figure 7.7 A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal ROUTE FINDING

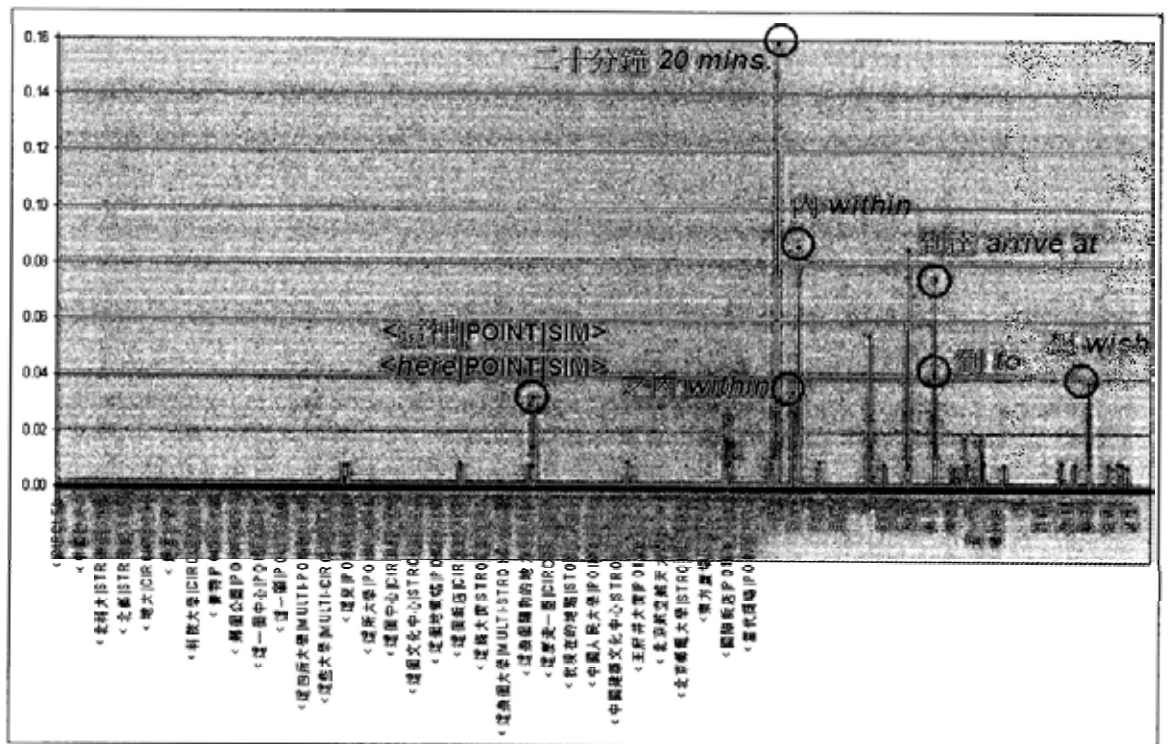


Figure 7.8: A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal TIME CONSTRAINT.

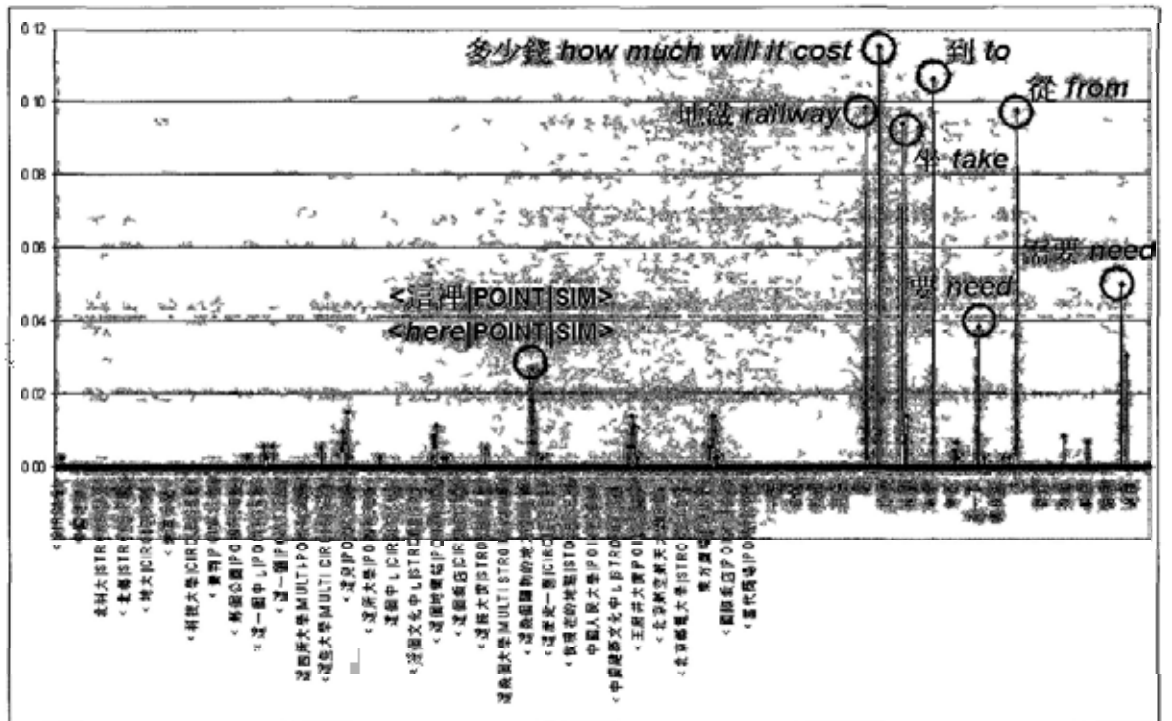


Figure 7.9 A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal TRANSPORTATION COSTS

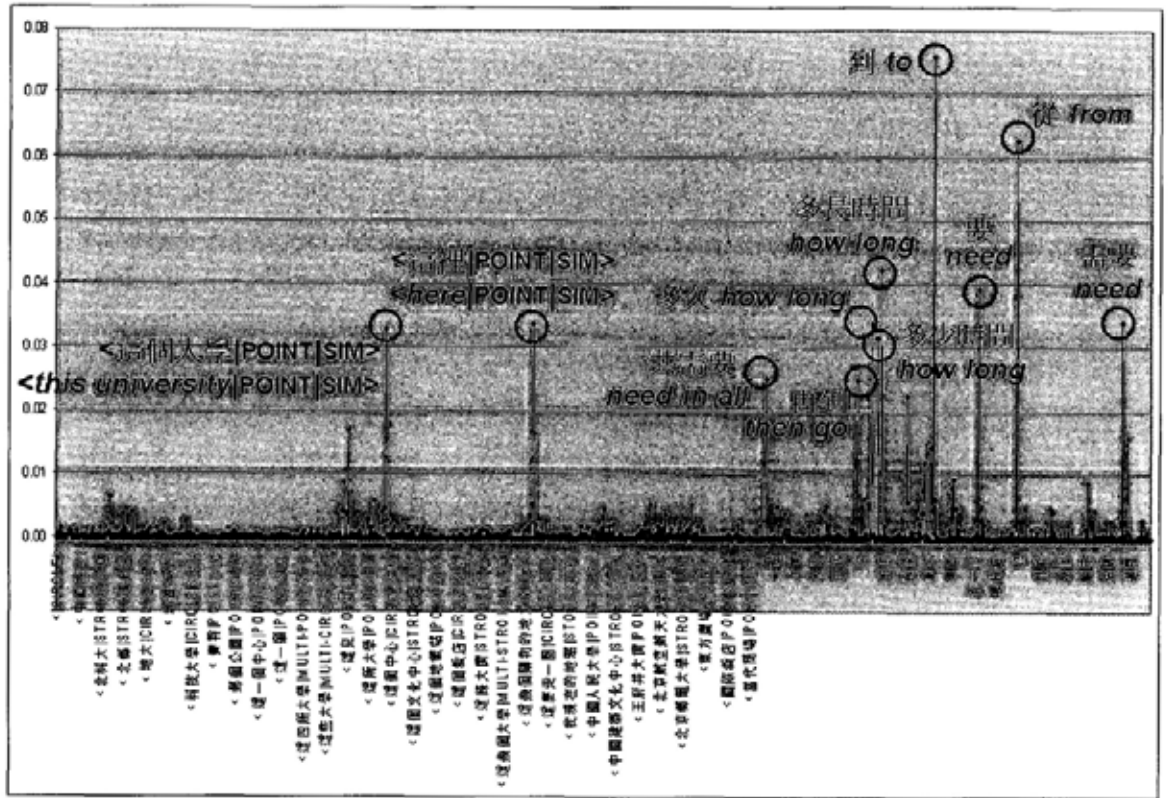


Figure 7.10: A plot of term weight from matrix  $\hat{B}$  against lexical and multimodal terms ( $M = 881$ ) for the task goal TRAVEL TIME.

## Chapter 8

# Latent Semantic Analysis for Task Goal Inference

This chapter describes our attempt in extending the semantic analysis framework presented in Chapter 7 for multimodal user input with speech and pen gestures on the associations between terms and inquiries. More specifically, our aim is to infer the domain-specific task goal(s) of the multimodal input with another formulation of the association matrix in latent semantic modeling (LSM). We can represent the associations between terms and inquiries as a term-inquiry matrix in LSM. This can be factorized into a term-semantics and an inquiry-semantics matrix using singular value decomposition (SVD). These two matrices associate terms and inquiries through an automatically derived space of semantics, instead of directly relating the terms with inquiries. We represent a multimodal input by means of lexical or multimodal terms. We then perform LSM to analyze the content of a multimodal input. Each input is associated with every latent semantic category by a weight. The weights are used for task goal inference. We would like to uncover the associations between terms and task goals through a data-derived latent space.

## 8.1 Latent Semantic Modeling for Task Goal Inference

In the previous Chapter, we apply latent semantic analysis to capture regularities in terms based on task goal. Similarly, we can apply latent semantic analysis for task goal inference based on multimodal input. LSM uses SVD to derive a latent semantic space that relates terms (combined lexical, gestural and multimodal terms) with the users' inputs. Correlations between cross-modal terms are captured from the training data. During testing, multimodal terms are extracted from the input and the vector is projected into the latent space. Thereafter, the task goal is inferred based on a combination of latent semantics.

### 8.1.1 Association Matrices

Associations between terms and inquiries can be summarized in a term-inquiry matrix  $G$ . Given  $M$  terms (details of the multimodal terms are presented in Section 4.6) and  $N$  inquiries, we form an  $M \times N$  matrix  $G$ . Each column represents an inquiry. The element  $g_{m,n}$ , is the weight (i.e. normalized term frequency using TF-IDF) for the term  $m$  in the  $n^{\text{th}}$  inquiry.

$$G = \begin{bmatrix} g_{1,1} & \cdots & g_{1,n} & \cdots & g_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{m,1} & \cdots & g_{m,n} & \cdots & g_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{M,1} & \cdots & g_{M,n} & \cdots & g_{M,N} \end{bmatrix} \quad (8.1)$$

where  $g_{m,n} = (1 - \varepsilon_m) \frac{\kappa_{m,n}}{\lambda_n}$ ,  
 $\varepsilon_m = -\frac{1}{\log N} \sum_{n=1}^N \frac{\kappa_{m,n}}{\tau_m} \log \frac{\kappa_{m,n}}{\tau_m}$ ,  
 $\frac{\kappa_{m,n}}{\lambda_n}$  is the term frequency of term  $m$ ,  
 $\log \frac{\kappa_{m,n}}{\tau_m}$  is the inverse document frequency of term  $m$ ,  
 $\kappa_{m,n}$  denotes the number of times the term  $m$  occurs in the  $n^{\text{th}}$  inquiry,  
 $\lambda_n$  is the total number of terms in the  $n^{\text{th}}$  inquiry,  
 $\varepsilon_m$  denotes the normalized entropy of term  $m$  in the data set; and  
 $\tau_m$  is the total number of times that term  $m$  occurs in the training set.

(8.2)

$G$  can be decomposed into a product of three matrices using SVD of order  $R$ , as shown in Equation 8.3.

$$\begin{aligned}
 G &= USV^T \\
 &= \begin{bmatrix} u_{1,1} & \cdots & u_{1,R} \\ \vdots & \ddots & \vdots \\ u_{M,1} & \cdots & u_{M,R} \end{bmatrix} \begin{bmatrix} s_{1,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_{R,R} \end{bmatrix} \begin{bmatrix} v_{1,1} & \cdots & v_{1,R} \\ \vdots & \ddots & \vdots \\ v_{N,1} & \cdots & v_{N,R} \end{bmatrix}^T \quad (8.3)
 \end{aligned}$$

where  $U$  is the term-semantics matrix of dimensions  $M \times R$ ,

$S$  is the diagonal matrix of singular values sorted in descending order with dimensions  $R \times R$ ,

$V$  is inquiry-semantics matrix of dimensions  $N \times R$ ,

$R = \min\{M, N\}$  is the order of decomposition and

$T$  is the transpose of the matrix.

$U$  and  $V$  are the left unitary matrix and right unitary matrix respectively. Each column of  $U$  contains the estimated weight of each term  $m$  that corresponds to the latent semantic category  $r$  while each column of  $V^T$  contains

the estimated weight of each inquiry  $n$  that corresponds to the latent semantic category  $r$ . Equation 8.3 projects the space of terms and inquiries onto a reduced  $R$ -dimensional space which is defined by the orthonormal basis given by the column vectors  $u_m$  and  $v_n$  from matrices  $U$  and  $V$  respectively. In order to collapse the terms that are “semantically similar”, we always choose  $R' < R$ . The smaller the value  $R'$ , the more pronounced is the reduction of semantic redundancy in the latent semantic space. Based on the latent semantic space, we may re-construct the space of terms and inquiries, denoted as  $\hat{G}$  in Equation 8.4.

$$G \approx \hat{G} = U\hat{S}V^T \quad (8.4)$$

where  $\hat{S}$  is the reduced diagonal matrix of singular values with optimized value of  $R'$  (i.e. with dimensions  $R' \times R'$ ).

We need to find an “optimal” choice of  $R'$  that minimizes semantic redundancy in the latent space, as well as minimizes the distortion between the re-constructed space  $\hat{G}$  and the original space  $G$ , in the implementation of Equation 8.4 in the training procedure. We plan to optimize  $R'$  through empirical analysis of the latent space.

### 8.1.2 Relating Task Goals with Latent Semantics

In the training procedure, we represent the  $n^{\text{th}}$  inquiry by the column vector  $g_n$  (Equation 8.5). The weights for latent semantic category  $r$  (i.e  $w_{n,r}$  as shown in Equation 8.7) can then be obtained by a dot product between  $g_n$  and the corresponding column vector of the left unitary matrix  $U$ ,  $u_r$  (Equation 8.6). Therefore, from the vector  $g_n$ , we can obtain a vector of weights  $w_n$  for each latent semantic category by Equation 8.8:



$$g_n = \begin{bmatrix} g_{1,n} \\ \vdots \\ g_{M,n} \end{bmatrix} \quad (8.5)$$

$$u_r = \begin{bmatrix} u_{1,r} \\ \vdots \\ u_{M,R'} \end{bmatrix} \quad (8.6)$$

$$w_{n,r} = g_n^T u_r \quad (8.7)$$

$$w_n = g_n^T U \quad (8.8)$$

where  $w_n = [w_{n,1} \ \dots \ w_{n,R}]$  and

$w_{n,r}$  is the weight of latent semantic category  $r$  for the  $n^{\text{th}}$  inquiry.

We use  $A$  to denote the total number of task goals within the application domain,  $a_n$  to denote the task goal of the  $n^{\text{th}}$  inquiry, and  $R'$  to denote the number of dimensions in the latent semantic space. We attempt to compute a projection matrix  $F$  that can transform the vector of weights for the latent semantic categories  $w_n$  into a vector of weights for the  $A$  task goals (see Equation 8.9).

$$h_n = w_n F \quad (8.9)$$

$$\text{where } F = \begin{bmatrix} f_{1,1} & \cdots & f_{1,A} \\ \vdots & \ddots & \vdots \\ f_{R',1} & \cdots & f_{R',A} \end{bmatrix},$$

$f_{r,a}$  is the weight of a latent semantic category  $r$  that would correspond to a task goal  $a$ ,

$$h_n = \begin{bmatrix} h_{n,1} & \cdots & h_{n,A} \end{bmatrix} \text{ and}$$

$h_{n,a}$  is the weight of the  $n^{\text{th}}$  inquiry would correspond to a task goal  $a$ .

According to Equation 8.9, associations between inquiry and latent semantic categories can be summarized in an inquiry-latent semantic categories matrix  $W$  (an  $N \times R'$  matrix) and the associations between inquiry and task goal can be summarized in an inquiry-task goal matrix  $H$  (an  $N \times A$  matrix). Therefore, we can obtain Equation 8.10 as follows.

$$H = WF \tag{8.10}$$

$$\text{where } W = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,R'} \\ \vdots & \ddots & \vdots \\ w_{N,1} & \cdots & w_{N,R'} \end{bmatrix} \text{ and}$$

$$H = \begin{bmatrix} h_1 \\ \vdots \\ h_N \end{bmatrix} = \begin{bmatrix} h_{1,1} & \cdots & h_{1,A} \\ \vdots & \ddots & \vdots \\ h_{N,1} & \cdots & h_{N,A} \end{bmatrix}.$$

Mathematically, the projection matrix  $F$  can be found using Equation 8.11.

$$F = W^{-1}H' \tag{8.11}$$

$$\text{where } H' = \begin{bmatrix} h'_1 \\ \vdots \\ h'_N \end{bmatrix} = \begin{bmatrix} h'_{1,1} & \cdots & h'_{1,A} \\ \vdots & \ddots & \vdots \\ h'_{N,1} & \cdots & h'_{N,A} \end{bmatrix},$$

$h'_n$  is a vector of manually labeled task goal for  $n^{\text{th}}$  inquiry,

$h'_{n,a}$  is the manually labeled task goal of inquiry  $n$ , in which

$$h'_{n,a} \in \{0, 1\} \text{ and } \sum_{a=1}^{a=A} h'_{n,a} = 1 \text{ and}$$

$W^{-1}$  is the pseudo inverse of the matrix  $W$ .

Through the projection matrix  $F$  and Equation 8.9, we can obtain the weight of each inquiry that would correspond to each task goal. A task goal  $a_n^*$  will be assigned as the automatic derived task goal for inquiry  $n$  where  $a_n^* = \arg \max_a \{h_{n,a}\}$ .

The performance of task goal inference of the training data can then be evaluated by comparing  $a_n^*$  to the manually annotated task goal  $a_n$ . Moreover, we may examine the structural relations between latent semantic category and task goals in the transformation matrix  $F$ .

In the testing procedure, we also represent the  $n^{\text{th}}$  inquiry by a vector  $g_n$ . We obtain the weights for the  $r$  latent semantic categories by Equation 8.8 where the left unitary matrix  $U$  is obtained from the training procedure. The vector of weights for each latent semantic category lies in the  $R'$ -dimensional space. We transform it to  $A$ -dimensional space and automatically derived task goal  $a_n^*$  for the  $n^{\text{th}}$  inquiry using Equation 8.9. The task goal inference performance can be evaluated by comparing the  $a_n^*$  assigned and task goal  $a_n$  manually annotated of the  $n^{\text{th}}$  inquiry. The TG inference accuracy is defined as:

$$\text{Accuracy} = \frac{\text{number of inquiries with correctly inferred TG}}{\text{Total number of inquiries in the data set}}$$

## 8.2 Task Goal Inference

In this section, we present the framework for inferring the task goal based on an input inquiry.

### 8.2.1 Performance Baseline using Vector-Space Model

As a reference baseline, we apply the vector-space model [82] for task goal inference as mentioned in Section 7.2.1. For each task goal  $a$ , we consider all of its training expressions and their multimodal terms. We create a vector  $j_a$  of weights, using the normalized term frequency TF-IDF of the multimodal terms. For an input multimodal expression, we create a vector  $g_n$ , similar to the column vector of  $G$  in Equation 8.1. The similarity between an inquiry  $g_n$  and task goal vector  $j_a$  is calculated as the inner product of the two vectors. Equation 8.12 shows the similarity calculation using the dot product between the unit vector of  $j_a$  and the unit vector of  $g_n$ .

$$similarity_{cosine}(j_a, g_n) = \frac{j_a \cdot g_n}{\|j_a\| \|g_n\|} \quad (8.12)$$

where  $j_a$  is the weight for all terms in the  $a^{th}$  task goal and  $g_n$  is the weight for all terms in the  $n^{th}$  inquiry.

The input expression is assigned to the task goal  $a_n^*$  which has the maximum similarity score, as shown in Equation 8.13.

$$a_n^* = \arg \max_a \{similarity_{cosine}(j_a, g_n)\} \quad (8.13)$$

Experiments show that vector-space model can correctly infer task goals for 84.5% (847/1002) and 82.5% (363/440) of the inquiries in training and test sets respectively.

### 8.2.2 Optimization of $R'$

Recall that the proposed approach using LSM involves setting up a term-inquiry matrix  $G$ . We include both lexical (unimodal, speech only) terms and multimodal terms with speech and pen gestures. There are a total of 314 unimodal terms and 567 multimodal terms in our training corpus. Hence the non-negative matrix  $G$  (in Equation 8.1) is of dimensions  $881 \times 1002$ . As described in Section 8.1, we apply SVD to  $G$  and factorize it into  $U$ ,  $S$  and  $V$ .

As mentioned before, the total number of lexical and multimodal terms sum to  $R = 881$ . We may consider that the original semantic space to be determined by these terms and attempt to determine the optimal number of dimensions for the latent space. We may choose the order of SVD approximation ( $R'$ ) with reference to the percentage of the cumulative sum of retained singular values over the maximum at  $R' = 881$ . We plot the percentage of the cumulative sum of preserved singular values over the total sum of all singular values (i.e. at  $R' = 881$ ). In Figure 8.1, we show the  $R'$  values corresponding to the cumulative sum of singular values, at multiples of 10%.

We also perform task goal inference on the multimodal inputs in the training set at different values of  $R'$  (see Figure 8.2). The performance of task goal inference increases with  $R'$ . The rate of increase slows down as  $R'$  becomes higher, reaching saturation approximately at  $R' = 309$  with a performance of task goal inference at 99.2% correct.

We also perform cross-validation of the performance of task goal inference in the training set at different values of  $R'$  between 235 and 309 (see Figure 8.3). The performance of task goal inference reaches saturation at  $R' = 263$ . The choice of  $R' = 263$  as the dimensionality of the latent space implies a reduction of 70% with respect to the original space of  $R = 881$ .

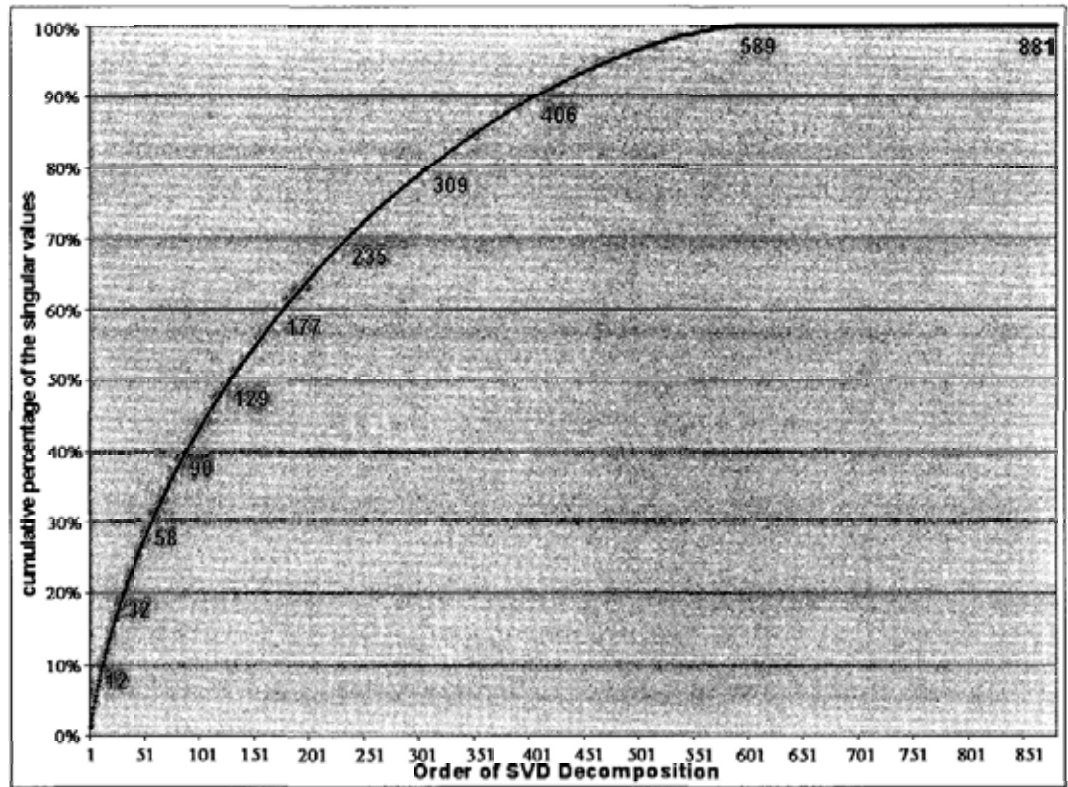


Figure 8.1: A plot of the cumulative percentage of the singular values against the order of SVD approximation.

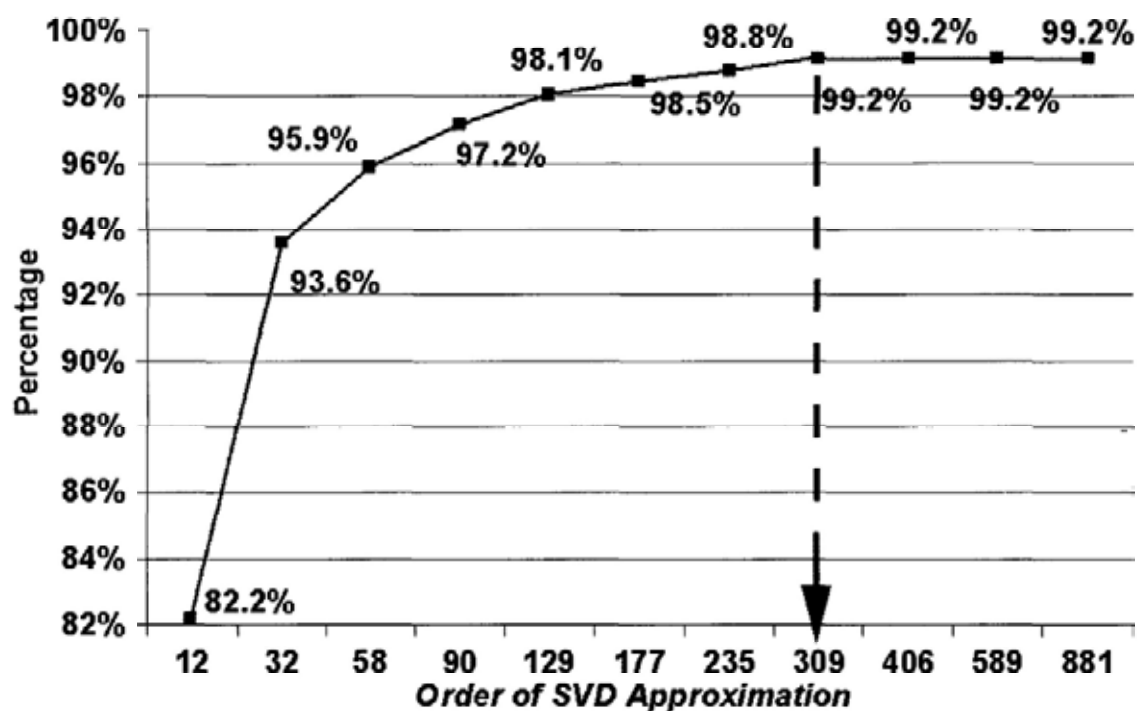


Figure 8.2: A plot of task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation.

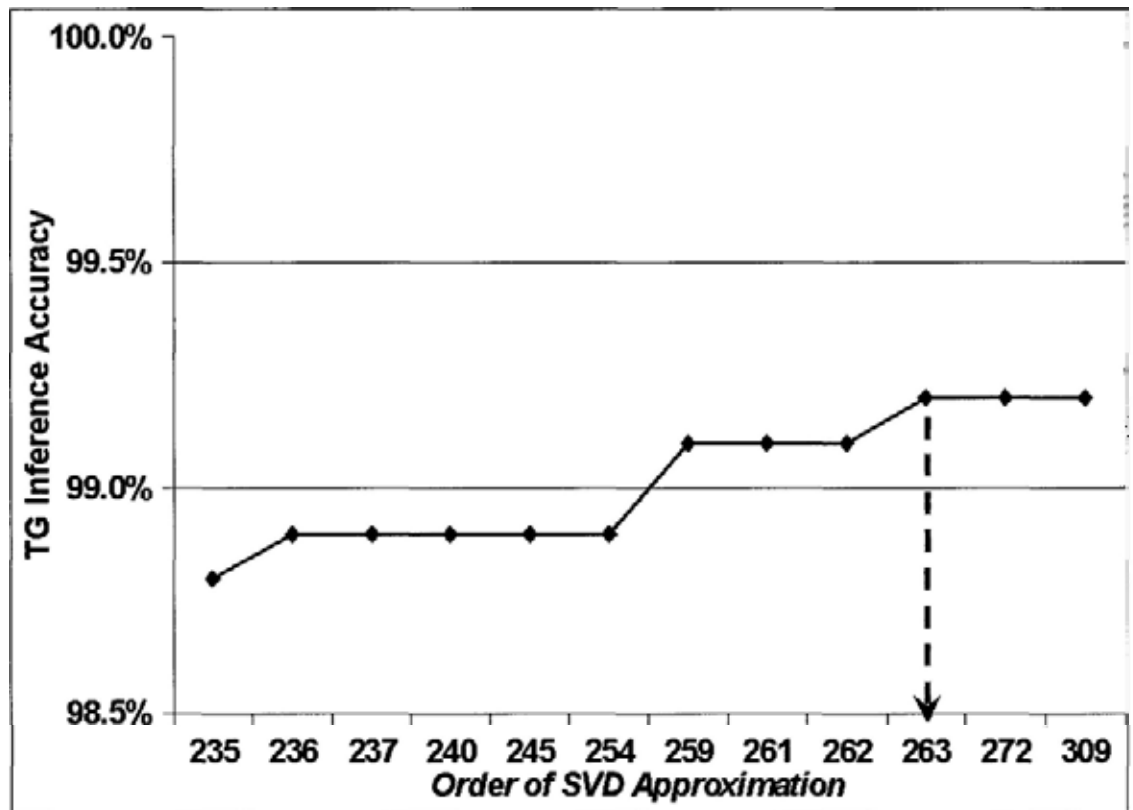


Figure 8.3: A plot of task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation for the optimization of  $R'$ .



### 8.2.3 Performance Evaluation

Overall performance in task goal inference for the training and test sets are 99.2% and 98.6% respectively. Detailed analyzes of the results are shown in Figure 8.4. The test set lacks inquiries that fall under the task goal of CHOICE OF VEHICLE (i.e., asking the user what type of vehicle he/she wishes to take). Performance of task goal inference remains high for all the other task goals (at 96% or above).

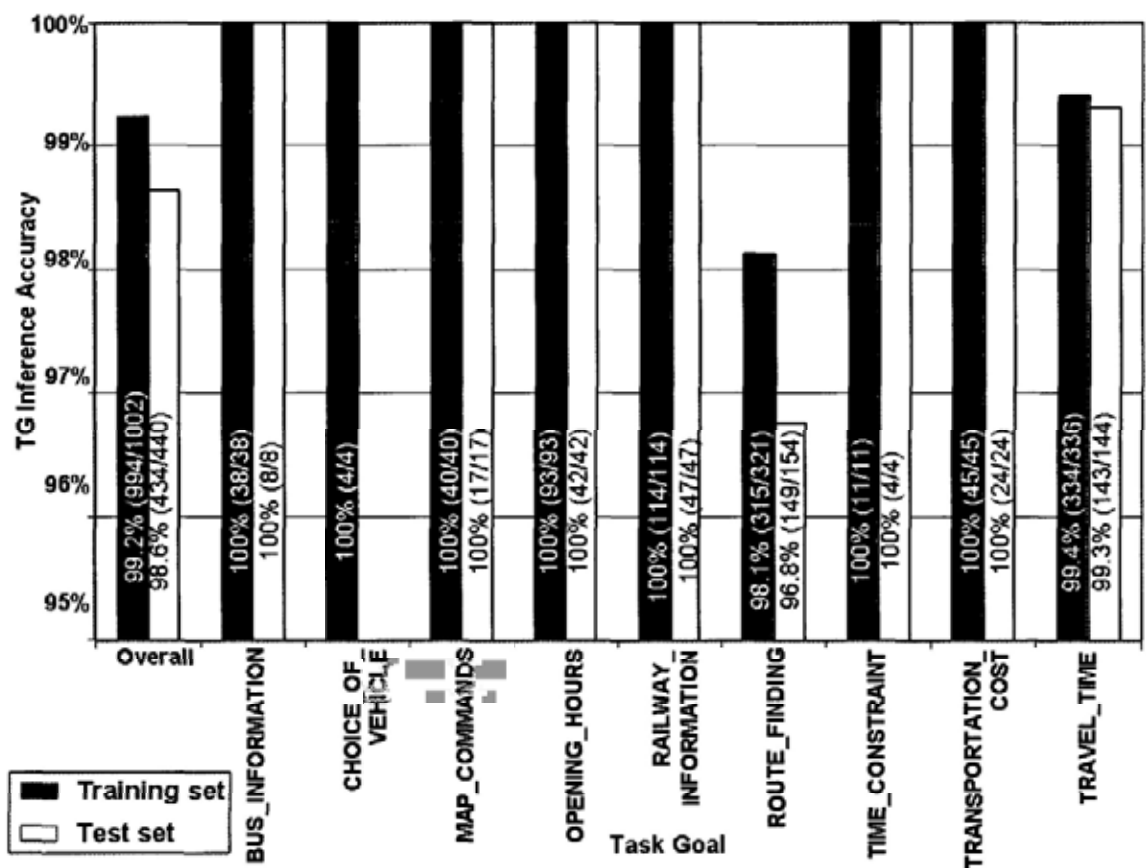


Figure 8.4: Performance of task goal inference for each of the nine task goals in the application domain. Results are based on the latent space with 406 dimensions.

## 8.3 Task Goal Inference with Spoken Terms Regularization

### 8.3.1 Spoken Terms Regularization

Analysis of the spoken inputs in Section 4.1 shows that there are many synonymous terms and aliases. For example, the word “route” in Chinese consists of two characters (i.e. 路線), which may also be reversed (as 線路) and the meaning of the word remains the same. Similarly, SLRs may have synonymous terms. For example, the full name 北京郵電大學 (i.e. *Beijing University of Post and Telecommunications*) may be abbreviated as 北郵 (i.e. *BUPT*). There is also a variety of verbalization to express the contextual phrase of “current location”, including: 目前的所在地, 當前的位置, 所在的地方, 所在地, etc. Other contextual phrases may differ by a “measure word” which is characteristic of Chinese, e.g., 這間大學 and 這所大學 both mean “this university”. In order to simplify processing, synonymous terms and aliases are collapsed into a single category. In other words, we have created a category for each group of semantically equivalent terms. It is conceivable that this categorization may be implemented through the use of SVD if sufficient data is available. Since we only have limited training data for the time being, we choose to design regularization rules (56 rules in all) for categorization.<sup>19</sup> As such, we have reduced the number of lexical terms significantly.<sup>20</sup> Since we also have pen gestures with their corresponding SLRs, we are still able to form “multimodal terms”. Each is a 3-tuple consisting of an SLR, the corresponding pen gesture and their temporal relationship as mentioned in Section 4.6.

---

<sup>19</sup>This step forms equivalence classes that group together terms with the same meaning. We expect that this step should help task goal inference because it reduces term diversity given the limited amount of training data.

<sup>20</sup>A lexical term refers to a tokenized Chinese word from the speech modality but which is not an SLR. Examples include 開放時間 *opening hours*, 路線 *route*, 從 *from*, etc.

	Before term regularization	After term regularization
Number of Multimodal terms	567	261
(SLR and pen)	508	233
(SLR only)	53	22
(Pen only)	6	6
Number of Lexical terms	314	216
Total number of terms	881	477

Table 8.1: Statistics of the lexical and multimodal terms (count by type).

The statistics of the lexical and multimodal terms in the training set are shown in Table 8.1. After regularization, the number of multimodal terms can be reduced to around 54.1% (477/881). The number of (SLR and pen) multimodal terms is fewer than expected. There are 22 multimodal terms that contain *only* an SLR with no pen gesture. This is because of an anaphoric reference (which can be resolved with contextual information). There are also 6 multimodal terms that contain pen gestures only and no SLR, due the use of ellipsis.

### 8.3.2 Performance Baseline using Vector-Space Model

As a reference baseline, we apply the vector-space model (see Section 8.2) for task goal inference. For each task goal  $a$ , we create a vector  $j_a$  of weights, using the normalized term frequency TF-IDF of the multimodal terms. For an input multimodal expression, we create a vector  $g_n$ , similar to the column vector of  $G$  in Equation 8.1. The similarity between an inquiry  $g_n$  and task goal vector  $j_a$  is calculated as the inner product of the two vectors with cosine normalization (see Equation 8.12). The input expression is assigned to the task goal  $a_n^*$  which has the maximum similarity score (see Equation 8.13).

Experiments show that vector-space model can correctly infer task goals for 90% (902/1002) and 87.5% (385/440) of the inquiries in training and test sets respectively. Table 8.3.2 shows the performance of task goal inference using vector-space model based on different weighting methods. Application of spoken terms regularization can reduce term diversity (especially reduce the term diversity between training and testing sets) and improve the task goal inference performance when compare the task goal inference performance obtained in Table 8.3.2 with the results presented in Table 7.2.1.

	Training set	Test set
Dot product (without cosine normalization) based on term frequency (see Equation 8.2)	79.2% (794/1002)	77.7% (342/440)
$similarity_{cosine}(j_a, g_n)$ (see Equation 8.12) based on term frequency	85.1% (853/1002)	85% (374/440)
Dot product (without cosine normalization) based on TF-IDF	87.8% (880/1002)	86.6% (381/440)
$similarity_{cosine}(j_a, g_n)$ based on TF-IDF	<b>90%</b> (902/1002)	<b>87.5%</b> (385/440)

Table 8.2: Task goal inference accuracy using the vector-space model approach based on different weight methods with spoken terms regularization.

### 8.3.3 Optimization of $R'$

Recall that the proposed approach using LSM involves setting up a term-inquiry matrix  $G$ . After spoken terms regularization, there are a total of 216 unimodal terms and 261 multimodal terms in our training corpus. Hence the non-negative matrix  $G$  (in Equation 8.1) is of dimensions  $477 \times 1002$ . We then apply SVD to  $G$  and factorize it into  $U$ ,  $S$  and  $V$ .

As mentioned before, the total number of lexical and multimodal terms sum to  $R = 477$ . We attempt to determine the optimal number of dimensions for the latent space (i.e. order of SVD approximation,  $R'$ ) with reference to the percentage of the cumulative sum of retained singular values over the maximum at  $R' = R = 477$ . We plot the percentage of the cumulative sum of preserved singular values over the total sum of all singular values (i.e. at  $R' = 477$ ). In Figure 8.5, we show the  $R'$  values corresponding to the cumulative sum of singular values, at multiples of 10%.

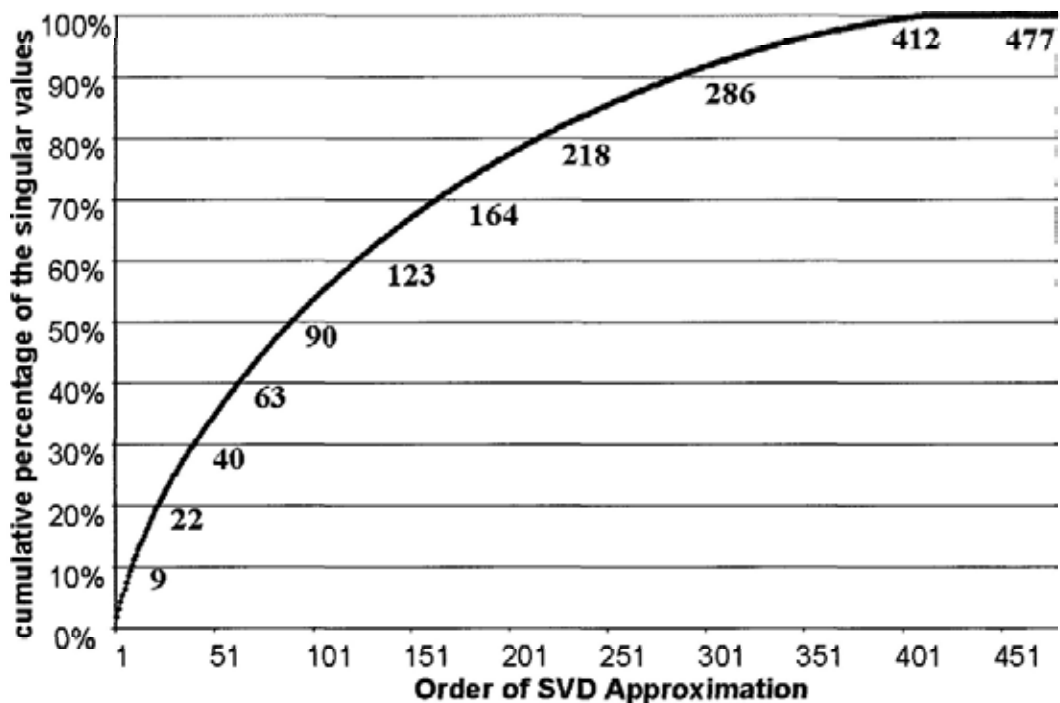


Figure 8.5: A plot of the cumulative percentage of the singular values against the order of SVD approximation.

We also perform task goal inference on the multimodal inputs in the training set at the different values of  $R'$  (see Figure 8.6). The performance reaches saturation approximately at  $R' = 286$ , with accurate task goal inference for 99.2% of the multimodal inputs.

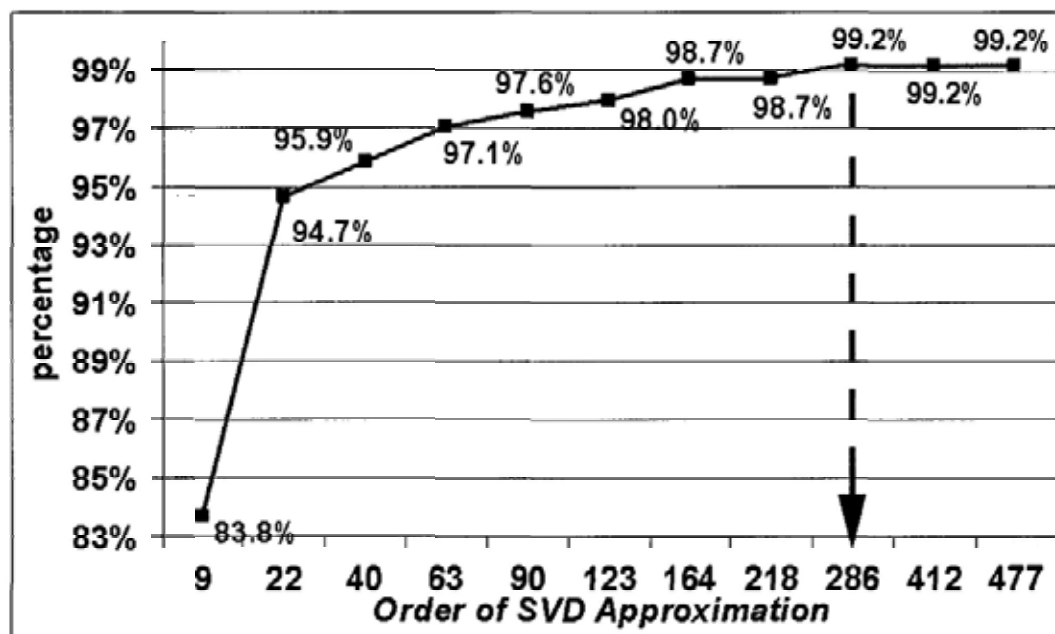


Figure 8.6: A plot of the task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation.

We also perform cross-validation of the performance of task goal inference in the training set at different values of  $R'$  between 218 and 286 (see Figure 8.7). The performance of task goal inference reaches saturation at  $R' = 263$ . The choice of  $R' = 263$  as the dimensionality of the latent space implies a reduction of around 45% with respect to the original space of  $R = 477$ .

### 8.3.4 Performance Evaluation

Overall performance in task goal inference for the training and test sets are 99.2% and 99.1% respectively. Table 8.3.4 shows the performance of task goal inference with and without spoken terms regularization. Detailed analysis of the results are shown in Figure 8.8. The test set lacks inquiries that fall under the task goal of CHOICE OF VEHICLE. Performance of task goal inference remains high for all the other task goals (at 98% or above).

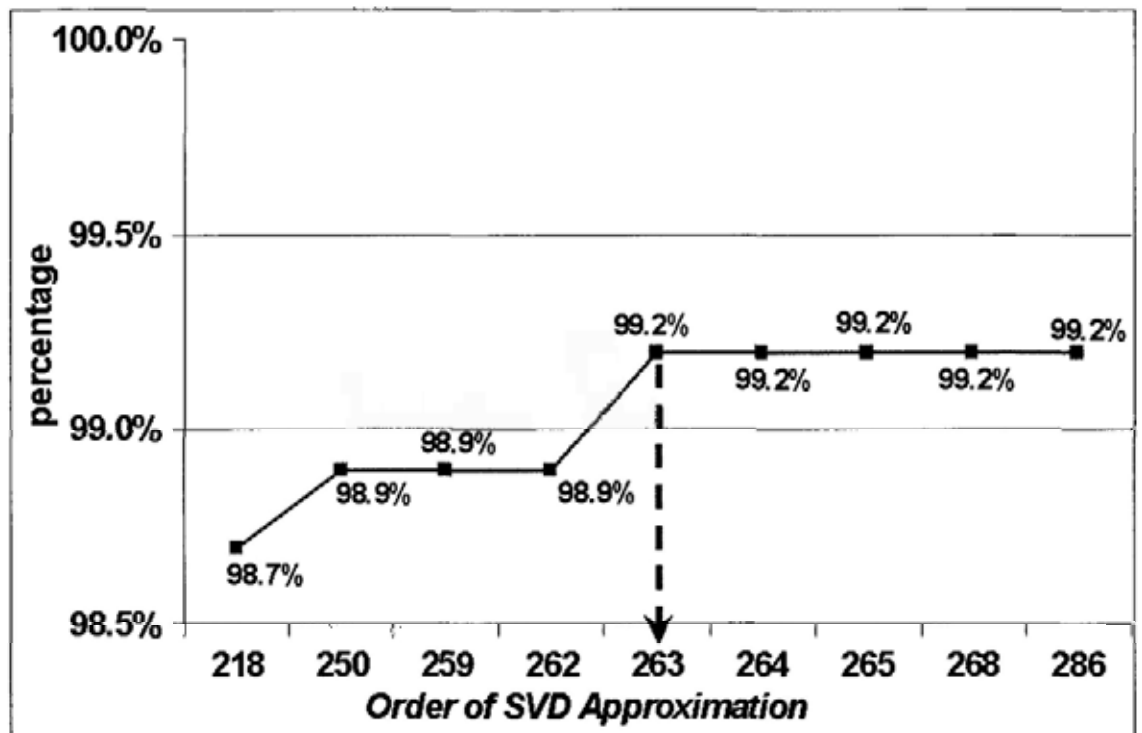


Figure 8.7: A plot of the task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation for the optimization of  $R'$ .

	Training set	Test Set
Performance without spoken terms regularization ( $M = 881$ and $R' = 263$ )	99.2%	98.6%
Performance with spoken terms regularization ( $M = 477$ and $R' = 263$ )	99.2%	99.1%

Table 8.3: Task goal inference accuracy before and after applying spoken terms regularization.

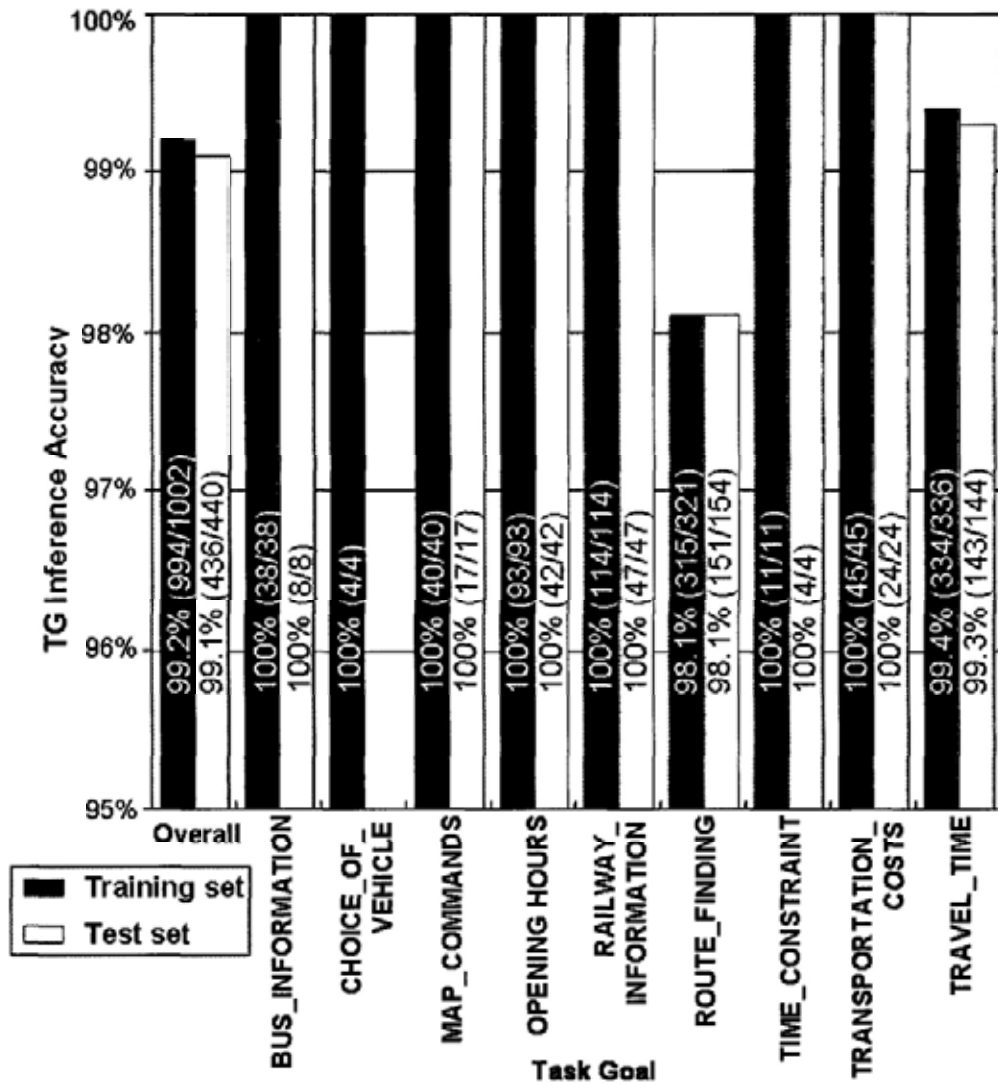


Figure 8.8: Performance of task goal inference for each of the nine task goals in the application domain. Results are based on the latent space with 263 dimensions.



## 8.4 Analysis of the Latent Semantic Space for Task Goal Inference

### 8.4.1 Sub-categorization of task goals

Analysis of the latent semantic space shows that it has sub-divided some of the task goals into logical sub-types. For example, the task goal `BUS INFORMATION` contains two latent semantic categories (see Figure 8.9 for the distribution):

- The latent semantic category ( $r = 13$ ) refers to `BUS INFORMATION` along a street;
- The category ( $r = 19$ ) refers to `BUS INFORMATION` within an area.

Table 8.4.1 shows the example of inquiries that belong to the latent semantic categories 13 and 19 in the training set for task goal `BUS INFORMATION`.



Figure 8.9: Percentage of multimodal inputs that belong to different latent semantic categories, within the task goal `BUS INFORMATION`. The numbers inside the bars are the labels (indexed by  $r$ ) of the latent semantic categories.

Another example is the task goal `OPENING HOURS`, which contains six latent semantic categories (see Figure 8.10 for their distribution):

- The latent semantic category ( $r = 11$ ) refers to `OPENING HOURS` of one location;

**Latent Semantic Category ( $r = 13$ ) – BUS INFORMATION along a street**

- 請問 行經 <這兒 | STROKE | SEQ> 的 公交車 路線

*What are the bus routes that pass through <here | STROKE | SEQ>?*

- 經過 <這裡 | POINT | SIM> 的 所有 公交 線路

*What are the bus routes that pass through <here | POINT | SIM>?*

- 列出 所有 經過 <這兒 | CIRCLE | SIM> 的 公共汽車

*List out all the bus routes that pass through <here | CIRCLE | SIM>?*

- 我 想 知道 經過 <這裡 | STROKE | SIM> 的 所有 的 公交路線

*I would like to know the bus routes that pass through <here | STROKE | SIM>?*

- 請 告訴 我 所有 經過 <這兒 | CIRCLE | SEQ> 的 公共汽車

*Please tell me the bus routes that pass through <here | CIRCLE | SEQ>?*

- 經過 <這條大街 | STROKE | SEQ> 的 所有 公交 線路 是 哪些

*What are the bus routes that pass through <this street | STROKE | SEQ>?*

- 從 <這條街 | STROKE | SEQ> 上 走 的 公車 有 哪些

*What are the bus routes that drive along <this street | STROKE | SEQ>?*

- 列出 所有 路過 <這條大街 | POINT | SIM> 的 公共汽車

*List out all the bus routes that pass through <this street | POINT | SIM>?*

- 我 想 知道 所有 的 行經 <這條大街 | STROKE | SIM> 的 公交路線

*I would like to know the bus routes that pass through <this street | STROKE | SIM>?*

- 告訴 我 所有 行經 <這條街 | STROKE | SIM> 的 公交路線

*Tell me the bus routes that pass through <this street | STROKE | SIM>?*

- 請 告訴 我 所有 公交車 經過 <這一條大街 | STROKE | SEQ> 的 號碼

*Please tell me the bus routes that pass through <this street | STROKE | SEQ>?*

- 我 想 知道 所有 路過 <建內大街 | CIRCLE | SEQ> 的 公共汽車

*I would like to know the bus routes that pass through <Jianguomen Inner Street | CIRCLE | SEQ>?*

<ul style="list-style-type: none"> <li>• &lt;建内大街   STROKE   SIM&gt; 都有 哪些 公交 <i>What are the bus routes that pass through &lt;Jianguomen Inner Street   STROKE   SIM&gt;?</i></li> <li>• 從 &lt;這邊   STROKE   SEQ&gt; 走的 公車 有 哪些 <i>What are the bus routes that pass through &lt;this side   STROKE   SEQ&gt;?</i></li> </ul>
<p><b>Latent Semantic Category (<math>r = 19</math>) – BUS INFORMATION within an area</b></p>
<ul style="list-style-type: none"> <li>• &lt;這兒   STROKE   SIM&gt; 附近 有 哪些 公交車 經過 <i>What are the bus routes that pass through the area around &lt;here   STROKE   SIM&gt;?</i></li> <li>• &lt;崇文門東大街   STROKE   SEQ&gt; 一百米 內 經過 的 公交車 有 哪些 <i>What are the bus routes that pass through the area of 100m from &lt;Chongwenmen East Street   STROKE   SEQ&gt;?</i></li> <li>• &lt;這條大街   STROKE   SEQ&gt; 一百米 內 經過 的 公交車 都有 哪些 <i>What are the bus routes that pass through the area of 100m from &lt;this street   STROKE   SEQ&gt;?</i></li> <li>• &lt;這兒   STROKE   SEQ&gt; 往 東 一百米 內 都有 哪些 公交 <i>What are the bus routes that pass through the area of 100m east from &lt;here   STROKE   SEQ&gt;?</i></li> <li>• 能否 告訴我在 &lt;這條街   CIRCLE   SEQ&gt; 一百米 內 的 公交車 路線 <i>Would you tell me the bus routes that pass through the area of 100m from &lt;this street   CIRCLE   SEQ&gt;?</i></li> <li>• 告訴我 所有 在 &lt;這個範圍   CIRCLE   SIM&gt; 行走 的 公交路線 <i>Tell me What are the bus routes that pass through &lt;this area   CIRCLE   SIM&gt;?</i></li> <li>• &lt;這兒   STROKE   SIM&gt; 附近 有 哪些 公交車 <i>What are the bus routes near the area around &lt;here   STROKE   SIM&gt;?</i></li> </ul>

Table 8.4: Examples of the inquiry that belong to the latent semantic categories 13 and 19 in the training set for task goal BUS INFORMATION.

- The category ( $r = 46$ ) refers to OPENING HOURS of single or multiple locations using ellipsis;
- The categories ( $r = 7$  and 29) refer to OPENING HOURS of multiple locations using multiple singular SLRs;
- The category ( $r = 9$ ) refers to OPENING HOURS of multiple locations using one aggregated SLR;
- The category ( $r = 12$ ) refers to OPENING HOURS of multiple locations using one plural SLR.

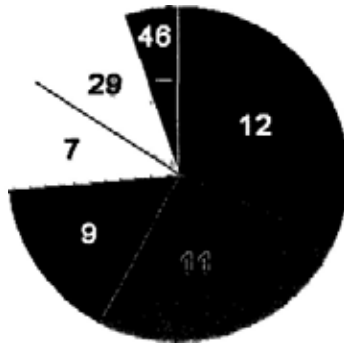


Figure 8.10: Percentage of multimodal inputs that belong to different latent semantic categories, within the task goal OPENING HOURS. The numbers inside the bars are the labels (indexed by  $r$ ) of the latent semantic categories.

We observe that latent semantic modeling has produced subcategories of specific task goals based on the ways in which users compose their inquiries. This is potentially advantageous because finer semantics categorization can enhance understanding and will facilitate automatic generation of system responses.

**Latent Semantic Category ( $r = 11$ ) – OPENING HOURS of one location**

- 請問 <這兒 | CIRCLE | SEQ> 什麼時候 開放  
*What is the opening hours of <here | CIRCLE | SEQ>?*
- 我 想 知 道 <這 裡 | POINT | SIM> 的 開 放 時 間  
*I want to know the opening hours of <here | POINT | SIM>.*
- <這兒 | POINT | SIM> 的 開 放 時 間  
*The opening hours of <here | POINT | SIM>.*
- <這 裡 | POINT | SEQ> 的 開 放 時 間 是 多 少  
*What is the opening hours of <here | POINT | SEQ>?*
- <這兒 | CIRCLE | SIM> 的 開 放 時 間 是 幾 點  
*What is the opening hours <here | CIRCLE | SIM>?*
- <這兒 | STROKE | SIM> 的 開 放 時 間 是 幾 點 到 幾 點  
*The opening hours of <here | STROKE | SIM>.*
- 我 想 知 道 <這 個 公 園 | POINT | SIM> 的 開 放 時 間  
*I would like to know the opening hours of <this park | POINT | SIM>.*
- <這 個 地 方 | POINT | SEQ> 什 麼 時 間 開 放  
*What is the opening hours of <this location | POINT | SEQ>?*
- 請 問 <雙 秀 公 園 | CIRCLE | SEQ> 何 時 開 放  
*What is the opening hours of <Shuangxiu Park | CIRCLE | SEQ>?*
- 我 想 知 道 <這 個 | POINT | SIM> 的 開 放 時 間  
*I would like to know the opening hours of <this | POINT | SIM>.*

**Latent Semantic Category ( $r = 46$ ) – OPENING HOURS of single or multiple locations using ellipsis**

- <  $\emptyset$  | POINT |  $\emptyset$  > 開 放 時 間  
*<  $\emptyset$  | POINT |  $\emptyset$  > Opening hours.*
- <  $\emptyset$  | MULTI-POINT |  $\emptyset$  > 營 運 時 間 分 別 是 什 麼  
*<  $\emptyset$  | MULTI-POINT |  $\emptyset$  > Opening hours of each.*

**Latent Semantic Category ( $r = 7$  and 29) – OPENING HOURS of multiple locations using multiple singular SLRs**

- 我想知道 <這個市場 | POINT | SIM> <這個廣場 | POINT | SIM> <這個購物中心 | POINT | SIM> 的營運時間

*I would like to know the opening hours of <this plaza | POINT | SIM>, <this plaza | POINT | SIM> and <this shopping center | POINT | SIM>.*

- 能告訴我 <這個市場 | POINT | SIM> <這個廣場 | POINT | SIM> <這個購物中心 | POINT | SIM> 的開放時間嗎

*Can you tell me the opening hours of <this plaza | POINT | SIM>, <this plaza | POINT | SIM> and <this shopping center | POINT | SIM>.*

- <這裡 | POINT | SIM> <這裡 | POINT | SIM> <這裡 | POINT | SIM> 的開放時間

*I would like to know the opening hours of <here | POINT | SIM>, <here | POINT | SIM> and <here | POINT | SIM>.*

- 我想查詢 <新東安市場 | POINT | SIM> <東方廣場 | POINT | SIM> <賽特購物中心 | POINT | SIM> 的運營時間

*I would like to enquire the opening hours of <Xindong'an Plaza | POINT | SIM>, <the Oriental Plaza | POINT | SIM> and <the Scitech Plaza | POINT | SIM>.*

- 列出 <新東安 | POINT | SIM> <東方廣場 | POINT | SIM> 還有 <賽特 | POINT | SIM> 的開放時間

*List out the opening hours of <Xindong'an | POINT | SIM>, <the Oriental Plaza | POINT | SIM> and <Scitech | POINT | SIM>.*

<p><b>Latent Semantic Category (<math>r = 9</math>) – OPENING HOURS of multiple locations using aggregated SLR</b></p> <ul style="list-style-type: none"> <li>● 勞駕 你 告訴我 &lt;這三個地方   MULTI-POINT   SIM&gt; 的 營業時間 <i>Please tell me the opening hours of &lt;these three places   MULTI-POINT   SIM&gt;.</i></li> <li>● &lt;這三個商場   MULTI-POINT   SEQ&gt; 什麼 時間 營業 <i>The opening hours of &lt;these three plazas   MULTI-POINT   SEQ&gt;.</i></li> <li>● 我 想 知道 &lt;這三個地方   MULTI-POINT   SEQ&gt; 的 營業時間 <i>I want to know the opening hours of &lt;these three places   MULTI-POINT   SEQ&gt;.</i></li> <li>● &lt;這三個商店   MULTI-POINT   SEQ&gt; 什麼 時候 上班 <i>What is the opening hours of &lt;these three shops   MULTI-POINT   SEQ&gt;.</i></li> <li>● &lt;這三個商場   MULTI-POINT   SIM&gt; 的 營業時間 是 <i>The opening hours of these &lt;three plazas are   MULTI-POINT   SIM&gt;?</i></li> <li>● 請問 &lt;這三個地方   MULTI-POINT   SEQ&gt; 什麼 時候 開放 <i>What are the opening hours of &lt;these three places   MULTI-POINT   SEQ&gt;?</i></li> </ul>
<p><b>Latent Semantic Category (<math>r = 12</math>) – OPENING HOURS of multiple locations using one plural SLR</b></p> <ul style="list-style-type: none"> <li>● 我 想 知道 &lt;這幾個地方   MULTI-POINT   SEQ&gt; 的 營運時間 <i>I would like to know the opening hours of &lt;these locations   MULTI-POINT   SEQ&gt;.</i></li> <li>● &lt;這些地方   MULTI-POINT   SEQ&gt; 的 營業時間 是 多少 <i>What are the opening hours of &lt;these locations   MULTI-POINT   SEQ&gt;?</i></li> <li>● &lt;這幾個購物的地方   MULTI-POINT   SEQ&gt; 的 營業時間 是 多少 <i>What are the opening hours of &lt;these shopping plazas   MULTI-POINT   SEQ&gt;?</i></li> <li>● 請問 &lt;這幾個地方   MULTI-CIRCLE   SEQ&gt; 的 開放時間 是 從 幾點 到 幾點 <i>The opening hours of &lt;these locations   MULTI-CIRCLE   SEQ&gt; are from when to when?</i></li> <li>● &lt;這幾個商場   MULTI-POINT   SIM&gt; 的 營運時間 是 幾點 到 幾點 <i>The opening hours of &lt;these plazas   MULTI-POINT   SIM&gt; are from when to when?</i></li> </ul>

Table 8.5: Examples of the inquiry that belong to the latent semantic categories 7, 9, 11, 12, 29 and 46 in the training set for task goal OPENING HOURS.

### 8.4.2 Capturing key terms for task goals

We examine the term weights in the latent semantic space to identify key terms that are indicative of each task goal. Illustrative examples include

- For the task goal MAP COMMANDS, key terms with the highest LSM weights are 放大 (i.e. *zoom in*), 縮小 (i.e. *zoom out*), 拉遠 (i.e. *zoom out*), as well as related standalone pen gestures expressed as the multimodal terms  $\langle \emptyset \mid \text{POINT} \mid \emptyset \rangle$  and  $\langle \emptyset \mid \text{CIRCLE} \mid \emptyset \rangle$
- For the task goal ROUTE FINDING, key terms with the highest LSM weights are 到 (i.e. *to*), 從 (i.e. *from*), 怎樣走 (i.e. *how to get to*), 最快 (i.e. *the fastest*), 依次 (i.e. *in sequence*), as well as the multimodal terms  $\langle \text{這裡} \mid \text{POINT} \mid \text{SEQ} \rangle$  (i.e.  $\langle \text{here} \mid \text{POINT} \mid \text{SEQ} \rangle$ ) and  $\langle \text{這個大學} \mid \text{POINT} \mid \text{SIM} \rangle$  (i.e.  $\langle \text{this university} \mid \text{POINT} \mid \text{SIM} \rangle$ )

Figures 8.11 and 8.12 are the plots of term weight from matrix  $\hat{G}$  against terms (both lexical and multimodal terms) for the task goal MAP COMMANDS and ROUTE FINDING respectively.

Moreover, the key terms identified through the latent semantic space between terms (both lexical and multimodal terms) and inquiries is consistent with the key terms identified through the latent semantic space between terms and task goals in Section 7.2.3.

### 8.4.3 Generalizing across related multimodal terms

Upon further examination of the LSM weights, we observe their ability to generalize across related multimodal terms, even if the correlations are not directly found in the training data. To describe the underlying mechanism - the LSM framework draws upon the co-occurrences between terms A and B, as well as the co-occurrences between B and C, in order to obtain the correlation between terms A and C.



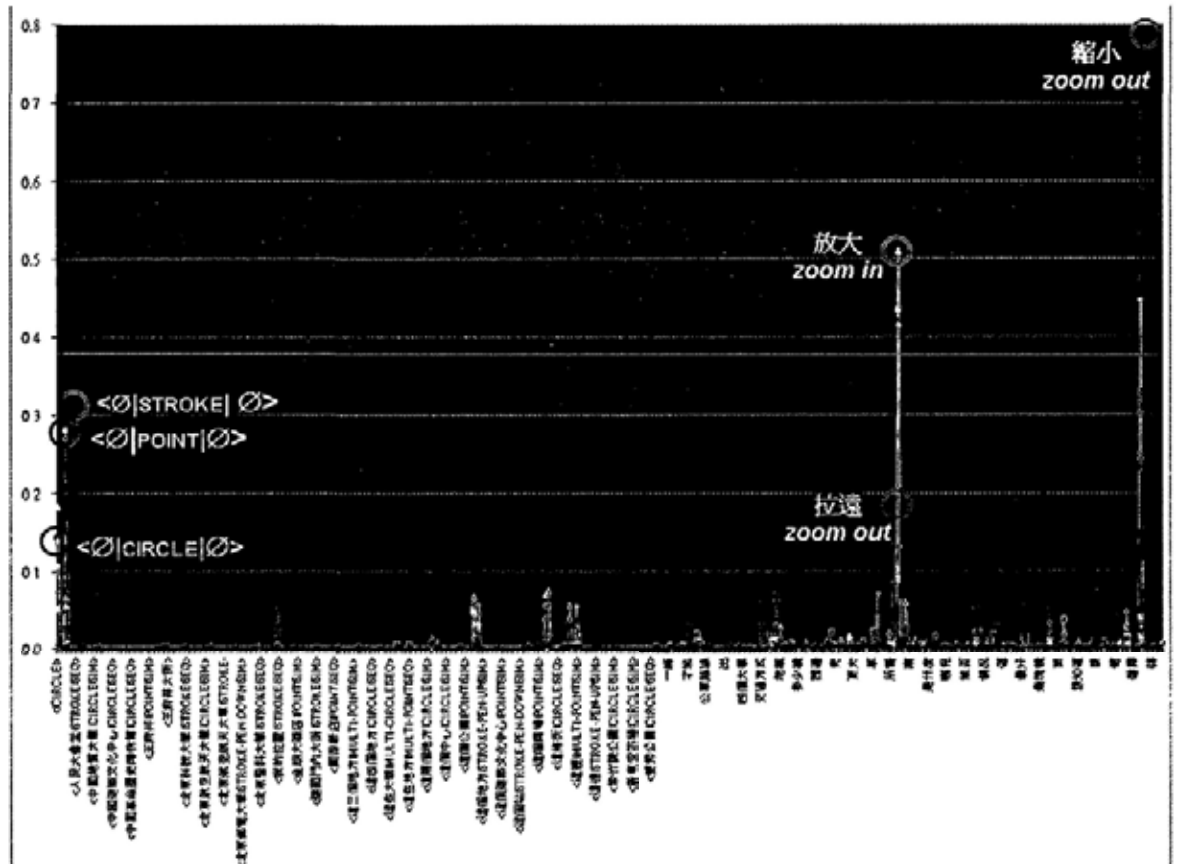


Figure 8.11: A plot of term weight from matrix  $\hat{G}$  against lexical and multimodal terms ( $M = 477$ ) for the task goal MAP COMMANDS.

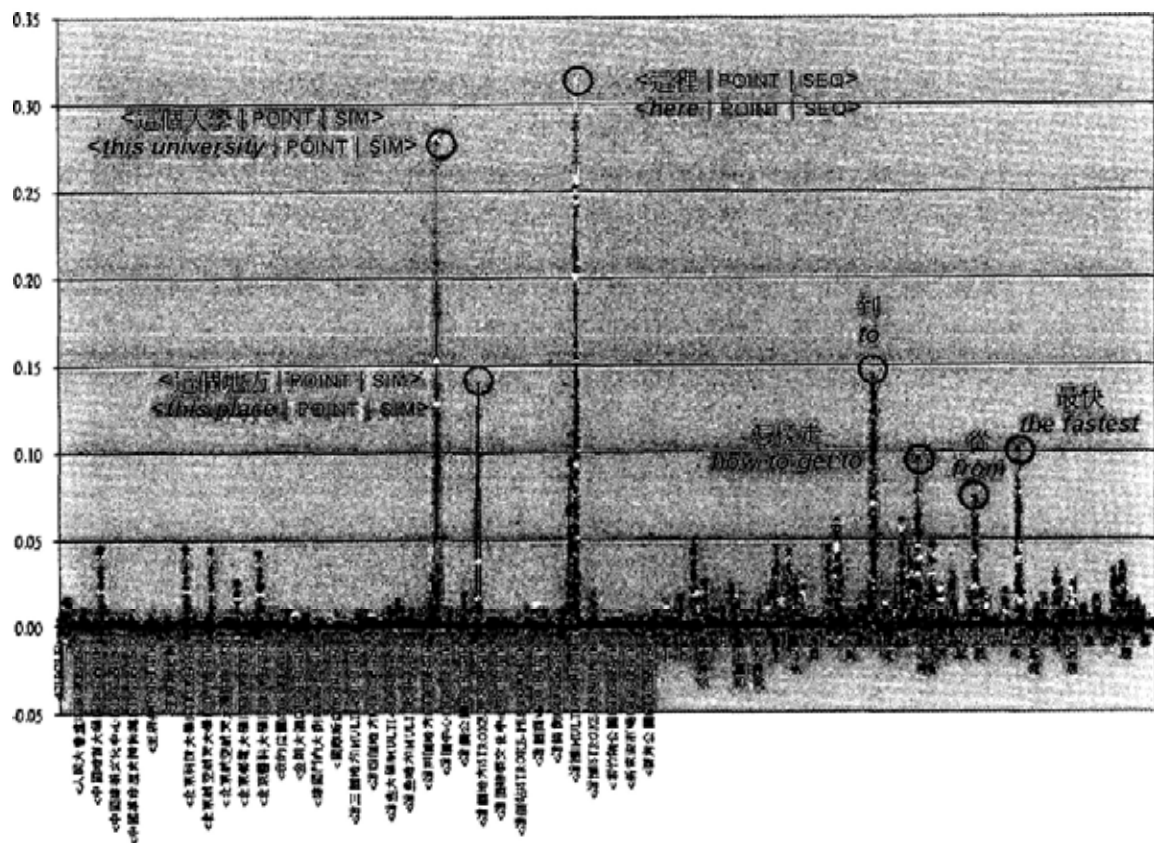


Figure 8.12: A plot of term weight from matrix  $\hat{G}$  against lexical and multimodal terms ( $M = 477$ ) for the task goal ROUTE FINDING.

As an illustration, we can refer to two multimodal inputs by which the user wishes to zoom in on a map

- 放大 CIRCLE (i.e. the verb phrase “zoom in” followed by a circle), corresponding respectively to the lexical and multimodal terms 放大 and  $\langle \emptyset \mid \text{CIRCLE} \mid \emptyset \rangle$
- 放大 POINT (i.e. the verb phrase “zoom in” followed by a point), corresponding respectively to the lexical and multimodal terms 放大 and  $\langle \emptyset \mid \text{POINT} \mid \emptyset \rangle$

The column vectors of these two input expressions, as extracted from the original term-inquiry matrix, are shown in Table 8.6. We compare these vectors with their counterparts in the reconstructed term-inquiry matrix  $\hat{G}$  (with  $R' = 263$ ), as shown in Table 8.7. We observe that the reconstructed column vector of the multimodal input “放大 CIRCLE” in Table 8.7 carry additional weighting ( $\geq 0.06$ ) for several additional multimodal terms, namely

- $\langle \text{這個地方} \mid \text{CIRCLE} \mid \text{SIM} \rangle$
- $\langle \text{這個範圍} \mid \text{CIRCLE} \mid \text{SEQ} \rangle$
- $\langle \text{這個範圍} \mid \text{CIRCLE} \mid \text{SIM} \rangle$  and
- $\langle \text{這幅圖} \mid \text{POINT} \mid \text{SIM} \rangle$

These additional multimodal terms with non-zero weights (see Table 8.7) did not appear in the original user inputs (see Table 8.6). But these terms are commonly used to convey the task goal MAP COMMAND, according to the training data (13 out of 40 multimodal inputs). LSM captures the new correlations among  $\langle \emptyset \mid \text{CIRCLE} \mid \emptyset \rangle$ , 放大 (i.e. *zoom in*),  $\langle \text{這個地方} \mid \text{CIRCLE} \mid \text{SIM} \rangle$  (i.e. *<this location | CIRCLE | SIM>*),  $\langle \text{這個範圍} \mid \text{CIRCLE} \mid \text{SEQ} \rangle$  (i.e. *<this area | CIRCLE | SEQ>*),  $\langle \text{這個範圍} \mid \text{CIRCLE} \mid \text{SIM} \rangle$  (i.e. *<this*

	放大 (i.e. <i>zoom in</i> ) < ∅   CIRCLE   ∅ >	放大 (i.e. <i>zoom in</i> ) < ∅   POINT   ∅ >
< ∅   CIRCLE   ∅ >	0.44	0
< ∅   POINT   ∅ >	0	0.34
放大 (i.e. <i>zoom in</i> )	0.37	0.37
<這個地方   CIRCLE   SEQ> < <i>this location</i>   CIRCLE   SEQ>	0	0
<這個地方   CIRCLE   SIM> < <i>this location</i>   CIRCLE   SIM>	0	0
<這個地方   POINT   SEQ> < <i>this location</i>   POINT   SEQ>	0	0
<這個地方   POINT   SIM> < <i>this location</i>   POINT   SIM>	0	0
<這個範圍   CIRCLE   SEQ> < <i>this area</i>   CIRCLE   SEQ>	0	0
<這個範圍   CIRCLE   SIM> < <i>this area</i>   CIRCLE   SIM>	0	0
<這個範圍   POINT   SIM> < <i>this area</i>   POINT   SIM>	0	0
<這個範圍   STROKE   SEQ> < <i>this area</i>   STROKE   SEQ>	0	0
<這幅圖   POINT   SIM> < <i>this map</i>   POINT   SIM>	0	0

Table 8.6: An excerpt of the term-inquiry matrix  $G$  corresponding to two multimodal inputs. The weights (shown up to 2 decimal places) are obtained using Equation 8.1. Translations are italicized.

	放大 (i.e. <i>zoom in</i> ) < ∅   CIRCLE   ∅ >	放大 (i.e. <i>zoom in</i> ) < ∅   POINT   ∅ >
< ∅   CIRCLE   ∅ >	0.18	<b>0.11</b>
< ∅   POINT   ∅ >	0.06	0.28
放大 (i.e. <i>zoom in</i> )	0.51	0.44
<這個地方   CIRCLE   SEQ> < <i>this location</i>   CIRCLE   SEQ>	0.00	0.00
<這個地方   CIRCLE   SIM> < <i>this location</i>   CIRCLE   SIM>	<b>0.07</b>	<b>0.05</b>
<這個地方   POINT   SEQ> < <i>this location</i>   POINT   SEQ>	0.00	0.00
<這個地方   POINT   SIM> < <i>this location</i>   POINT   SIM>	0.03	<b>0.05</b>
<這個範圍   CIRCLE   SEQ> < <i>this area</i>   CIRCLE   SEQ>	<b>0.07</b>	0.04
<這個範圍   CIRCLE   SIM> < <i>this area</i>   CIRCLE   SIM>	<b>0.07</b>	0.04
<這個範圍   POINT   SIM> < <i>this area</i>   POINT   SIM>	0.00	0.00
<這個範圍   STROKE   SEQ> < <i>this area</i>   STROKE   SEQ>	0.00	0.00
<這幅圖   POINT   SIM> < <i>this map</i>   POINT   SIM>	<b>0.06</b>	<b>0.06</b>

Table 8.7: An excerpt of the reconstructed term-inquiry matrix  $\hat{G}$  corresponding to two multimodal inputs as in Table 8.6. The estimated weights (shown up to 2 decimal places) of  $\hat{G}$  are obtained using Equation 8.4 with  $R' = 263$ . Translations are italicized.

*area* | CIRCLE | SIM> and <這幅圖 | POINT | SIM> (i.e. <*this map* | POINT | SIM> and put them into correlated latent semantics. The weights in Table 8.7 reflect that the circling action can be used to indicate a single location (i.e. 這個地方) or an area (i.e. 這個範圍).

Similarly, we also observe that the feature vector of the multimodal input “放大 POINT” in Table 8.7 introduces additional multimodal terms with non-zero weights (e.g. 0.05) for several additional multimodal terms:

- <這個地方 | CIRCLE | SIM>
- <這個地方 | POINT | SIM> and
- <這幅圖 | POINT | SIM>

These additional multimodal terms with non-zero weights (see Table 8.7) did not appear in the original user inputs (see Table 8.6). But these terms are commonly used to convey the task goal MAP COMMAND (11 out of 40 multimodal inputs). LSM captures the new correlations among < $\emptyset$  | POINT |  $\emptyset$ >, 放大 (i.e. *zoom in*), <這個地方 | CIRCLE | SIM> (i.e. <*this location* | CIRCLE | SIM>), <這個地方 | POINT | SIM> (i.e. <*this location* | POINT | SIM>) and <這幅圖 | POINT | SIM> (i.e. <*this map* | POINT | SIM>) and put them into correlated latent semantics. The weights in Table 8.7 reflect that the pointing action can be used to indicate a single location (i.e. 這個地方).

## 8.5 Error Analysis of the Latent Semantic Space for Task Goal Inference

Error analysis of the latent semantic space shows that task goal inference errors is mainly due to the ambiguity of a specific time expression (e.g. 二十分鐘, i.e. “*20 minutes*”).

During data collection, one of the tasks is 告知系統你要在二十分鐘內到達國際飯店 (i.e. *tell the system that you have to arrive at the International Hotel in 20 minutes*). Since we did not randomize the duration of the time phrase 二十分鐘 (i.e. *20 minutes*), it has become a key term of the task goal TIME CONSTRAINT with relatively high term weight. Therefore, whenever an inquiry contains the time phrase 二十分鐘 (i.e. *20 minutes*), it has been inferred as the task goal TIME CONSTRAINT. In order to prevent the same problem, we should randomize all the numeric values (including time expressions) in the future data collection. Table 8.8 shows the example of inquiries that belong to the task goal ROUTE FINDING but incorrectly infer as the task goal TIME CONSTRAINT.

## 8.6 Chapter Summary

In this chapter, we have extended our study to the usage pattern analysis and automatic task goal inference of multimodal user inputs with speech and pen gestures. We use a non-negative term-inquiry matrix to capture the associations between terms (lexical and multimodal) and inquiries. Decomposition of the term-inquiry matrix using singular value decomposition captures the associations between terms and inquiries through a latent semantic space. We project the latent semantic space into the space of task goals through a matrix derived from training data. An input multimodal inquiry can be projected into the latent semantic space and then into the task goal space. This gives a vector with which we can use the highest weighting element to select the inferred task goal. We experimented with this approach based on the manually transcribed multimodal corpus. Analysis shows structural relations between latent semantic categories for certain task goals. Furthermore, the weights of the lexical and multimodal terms in the latent semantic space can also help us identify key terms for specific task goals. The latent semantic approach

- 怎樣 在 二十分鐘 內 到達 <這個飯店 | POINT | SIM>  
*How can I to go to <this hotel | POINT | SIM> within 20 minutes?*
- 怎麼樣 在 二十分鐘 內 到達 <國際飯店 | CIRCLE | SEQ>  
*How can I to go to <the International Hotel | CIRCLE | SEQ> within 20 minutes.*
- 如何 在 二十分鐘 內 到達 <國際飯店 | POINT | SIM>  
*How can I to go to <the International Hotel | POINT | SIM> within 20 minutes?*
- 二十分鐘 內 到 <國際飯店 | POINT | SIM> 怎麼走  
*Go to <the International Hotel | POINT | SIM> within 20 minutes. How can I go?*
- 我在 <這兒 | CIRCLE | SIM> 想 二十分鐘 到 <國際飯店 | ◊ | ◊ > 怎麼走  
*I'm now at <here | CIRCLE | SIM>. Want to go to <the International Hotel | ◊ | ◊ > within 20 minutes, how can I go?*
- 請 給出 一條 能 在 二十分鐘 內 到達 <國際飯店 | CIRCLE | SEQ> 的路線  
*Please suggest a route that can arrive at <the International Hotel | CIRCLE | SEQ> within 20 minutes?*

Table 8.8: Examples of inquiries that belong to the task goal of ROUTE FINDING but are incorrectly infer as TIME CONSTRAINT. Translations are italicized.



achieves around 99% accuracy in task goal inference, for both the training and test sets. This is significantly higher than the reference baseline obtained with a vector-space model, which achieves 90% and 87.4% for the training and test sets respectively.

## Chapter 9

# Conclusions and Future Work

### 9.1 Thesis Summary

This thesis explored the cross-modality semantic integration method with hypothesis rescoring for robust interpretation in multimodal interface. Correct cross-modality semantic integration enables our framework to the multimodal input expression to be paraphrased as a unimodal (speech-only) input, for subsequent processing of our existing spoken dialog system [72] [85] [86] with dialog and discourse modeling and natural language generation components. Hence the cross-modality semantic integration framework offers an elegant front-end extension to our dialog system, to enable it to handle both unimodal (speech-only) as well as multimodal (speech and pen) inputs.

In order to support our investigations in:

- characterization and extraction of features from speech and pen modalities;
- recognition of input events from each modality (i.e. spoken locative references in speech and pen gestures in pen input);
- interpretation of the recognition output of spoken locative references and pen gestures as their partial semantics;

- integration of the partial semantics across modalities;
- maintaining robustness against imperfectly captured inputs and misrecognitions; and
- interpretation of the user's intention by integration across the multiple modalities,

we designed and collected a multimodal corpus in domain of city navigation around the Beijing area. This corpus contains 1,518 multimodal expressions with frequent locative references. The speech and pen modalities have been transcribed by hand. The inquiries cover nine pre-defined task goals. The task goal of each multimodal input is hand-labeled as a gold standard. We have also manually annotated the domain-specific named entities and SLRs in the transcribed speech and manually annotated the cross-modality pairings between an SLR from speech and a pen gesture. An SLR may map to zero, one or multiple pen gesture(s) and vice versa. We begin with an analysis of the usage patterns and designed the format of a multimodal term to be a 3-tuple, consisting of an SLR, pen gesture(s) and their temporal relationship (i.e. <SLR | pen\_gesture\_type | temporal\_relationship>). Such multimodal terms can represent the cross-modality integration patterns adopted by the user. Characteristic cross-modal integration patterns are derived from the training set to form multimodal terms. We also derive lexical terms from the speech portion of the multimodal expression. Processing of the speech and pen input modalities with automatic speech and pen recognition components shows that the overall Mandarin speech character recognition and pen gesture type recognition accuracies are 44.6% and 86.6% respectively.

After characterization of the multimodal input with speech and pen gestures, we present a framework pertaining to automatic semantic integration of the multimodal inputs. The two input modalities (speech and pen) abstract the user's intended message differently into input events (i.e. key

terms/phrases in speech and different gestures in the pen modality). The semantics of an input event may be imprecise (e.g. a pen stroke on a map may denote a street or demarcation), incomplete (e.g. use of anaphora in “*how about the previous one?*”) or erroneous due to mis-recognitions. The proposed framework begins by generating (partial) interpretations for each input event, which are represented as a ranked list of hypothesized interpretations. We devise a cross-modality semantic integration procedure to align input events in the speech modality with those in the pen modality using the Viterbi alignment algorithm [62]. Cost functions are designed to enforce the constraints of temporal ordering of the input events in each modality, as well as the semantic compatibility between hypothesized interpretations across modalities. Hence the alignment integrates across modalities and disambiguates among possible interpretation alternatives to decode the user’s holistic communicative intent. Application of cross-modality integration to these near-perfect transcripts (i.e. manual transcription) generated correct unimodal paraphrases for over 97% of the training and testing sets. However, if we replace this with the top-scoring speech and top-scoring pen recognition transcripts (which contain errors), the performance drops to 52% for both training and test sets. Analysis shows that complementarity and redundant relations between SLRs and pen gestures can salvage the performance of cross-modality integration in the presence of recognition errors through mutual disambiguation and mutual reinforcement [17].

In order to achieve robustness towards imperfect transcripts, we extend our framework with a hypothesis rescoring procedure. For each multimodal expression, this procedure considers all candidates for cross-modality integration based on the  $N$ -best ( $N = 100$ ) speech recognition hypotheses and the  $M$ -best ( $M = 4$ ) pen input recognition hypotheses. Note that a recognized pen gesture can generate  $Q$  location hypotheses that are fed into the cross-modality hypothesis rescoring procedure. Rescoring combines such elements

as the integration scores obtained from the Viterbi algorithm,  $N$ -best purity for recognized spoken locative references (SLRs), as well as distances between coordinates of recognized pen gestures and relevant icons on the map. Experiments using the  $N$ -best ( $N = 100$ ) speech recognition hypothesis and  $M$ -best ( $M = 4$ ) pen recognition hypotheses show that the rescoring and re-ranking helped improve the performance of correct cross-modality interpretation significantly to 71.8% and 72.7% for the training and testing sets, respectively.

In order to analyze the correlations between the two modalities, we have also performed a comparative analysis of manually transcribed multimodal ( $MM$ ) user inputs together with their automatically generated, semantically equivalent unimodal ( $UM$ ) counterparts. These are generated by the cross-modality framework proposed. We trained a class trigram language model with 1,450 multimodal and unimodal speech utterances and compared the perplexities ( $PP$ ) between parallel multimodal and unimodal test sets (with 430 utterances each). We observe that the speech components of multimodal expressions are generally shorter with lower lexical variability than their unimodal counterparts. Comparison with per-utterance perplexities affirms the relationships of complementarity and redundancy across the speech and pen modalities. One subset of our data exhibits the equality of ( $PP_{MM} = PP_{UM}$ ) and consists mainly of multimodal expressions where speech and pen modalities carry redundant semantics. The other subset exhibits the inequality of ( $PP_{MM} < PP_{UM}$ ) where the speech and pen modalities carry complementary semantics. We also observe the occurrences of ellipses, where certain semantics appear in one modality but not the other, and form a special case of complementarity. These observations have implications on the choice of fusion architectures for multimodal input interpretation.

On the interpretation of the multimodal inputs, we have applied latent semantic analysis for task goal inference of the multimodal inputs. We use a

non-negative term-inquiry matrix to capture the associations between terms (lexical and multimodal) and inquiries. Decomposition of the term-inquiry matrix using singular value decomposition captures the associations between terms and inquiries through a latent semantic space. We project the latent semantic space into the space of task goals through a matrix derived from training data. An input multimodal inquiry can be projected into the latent semantic space and then into the task goal space. This gives a vector with which we can use the highest weighting element to select the inferred task goal. We experimented with this approach based on the multimodal corpus. Analysis shows structural relations between latent semantic categories for certain task goals. Furthermore, the weights of the lexical and multimodal terms in the latent semantic space can also help us identify key terms for specific task goals. The latent semantic approach achieves around 99% accuracy in task goal inference, for both the training and test sets. This is significantly higher than the reference baseline obtained with a vector-space model, which achieves 90% and 87.4% for the training and test sets respectively.

## 9.2 Contributions

The contributions of this thesis can be summarized as follows:

- **Propose a cross-modality semantic integration framework for robust interpretation of multimodal input with speech and pen gestures**

The framework begins by generating partial interpretation of each modality and integrating them by the Viterbi alignment algorithm by incorporating temporal order and semantic compatibility constraints. The framework then considers the ranked confidence of multiple recognition hypotheses in both modalities based on the elements of  $N$ -best purity of speech, distance measure of pen gesture and their integration score.

The framework is able to handle multiple multimodal input events in a complex input expression (e.g. a navigational inquiry that involves a composition of singular, plural and aggregated locative references).

- **Implement a prototype of the framework**

A prototype of the cross-modality semantic integration framework on the task of navigation around Beijing is implemented. The client-side interface of the prototype is developed on a Pocket PC, which is used to show the map of Beijing and results of integration. It also contains a home-grown pen gesture recognizer. The server-side of the prototype is developed on a notebook PC which consists of a Mandarin character recognizer [67], speech and pen gesture interpretation components, integration components and the hypothesis rescoring component. Once a user makes a multimodal input on the Pocket PC on the client-side, the recorded speech and the recognized pen gesture information are transmitted to the server through socket for processing. The final result will be sent back to the client-side Pocket PC when ready.

- **Investigate the relationships of complementarity and redundancy across modalities**

We have performed a comparative analysis between multimodal inputs with their corresponding semantically equivalent unimodal paraphrases. The unimodal paraphrases are generated by the cross-modality semantic integration framework mentioned before using Viterbi alignment algorithm. We trained a class trigram language model and compare the per-utterance perplexities (PP) between parallel multimodal (MM) and unimodal (UM) test sets. Comparison affirms the relationships of complementarity ( $PP_{MM} < PP_{UM}$ ) and redundancy ( $PP_{MM} = PP_{UM}$ ) across speech and pen modalities. We also observe the occurrences of ellipses, where certain semantics appear in one modality but not the other, and

forms a special case of complementarity.

- **Apply latent semantic modeling for task goal inference of multimodal user inputs**

We use a non-negative term-inquiry matrix to capture the associations between terms (lexical and multimodal) and inquiries. Decomposition of the term-inquiry matrix using singular value decomposition captures the associations between terms and inquiries through a latent semantic space. An input multimodal inquiry can be projected into the latent semantic space and then into the task goal space for the selection of inferred task goal through a matrix derived from training data. Analysis of the latent semantic space shows structural relations between latent semantic categories for certain task goals. Furthermore, the weights of the lexical and multimodal terms in the latent semantic space can also help us identify key terms for specific task goals.

### 9.3 Future Work

According to the error analysis, the majority of errors that are due to the presence of redundant SLR(s) can be solved by incorporating timing information. In the future, attention should be paid to detection of a user's integration pattern. Whenever a user is detected as a simultaneous integrator, timing information (i.e. temporal difference between a SLR and a pen gesture) should be incorporated into the cost function of the Viterbi alignment algorithm. However, it can only be applied to simultaneous integrators but not sequential integrators since the integration pattern of a sequential integrator is mainly based on temporal order.

Analysis on the use of ellipsis in multimodal input is also suggested as an area for future work. Correlation between SLR(s) and pen gesture(s) does not



exist in the multimodal input with the use of ellipsis (due to the omission of SLR(s) in the multimodal input). A pen gesture input is ambiguous to the multimodal system (e.g. is a “drag-and-drop” action recognized as a stroke or a map movement?). Currently, we use the “click-to-speak” method to handle the case where all pen actions occur after the click of the “start” button, such that they will be considered as part of the multimodal input and otherwise considered as map movement.

Moreover, since our current framework is focused on the alignment between SLR(s) and pen gesture(s), integration methods between semantics from speech and pen modalities in the presence of ellipsis (i.e. in the absence of SLR) is also challenging as we do not know which semantics should be integrated. Analysis of the syntactic structure of multimodal input may be useful for the handling ellipsis.

A possible direction is the analysis of the phonological peak of the spoken input. As mentioned in McNeill [1], pen gestures are integrated into the phonology of the spoken input. Chen [39] also showed that there is correlation between the delta pitch value in the speech signal and occurrence of deictic-like gestures. It is possible for us to analyze the “peak” of the spoken input and investigate the possibility of integrating a pen gesture into those “peaks” even though an SLR cannot be recognized during that duration of spoken input. We may analyze the recognition errors of the spoken input and generate a confusion matrix for the expansion on phonetic confusion during speech recognition, especially for the phone(s) that is/are at the peak(s) of the spoken input.

On the extension of a pen gesture recognizer, it is suggested to support more pen gesture types (e.g. arrow and different types of encircling). Moreover, our current pen gesture recognizer can automatically filter out redundant pen gestures of the same pen gesture type based on the difference in time and

distance. The ability to filter redundant pen gestures in different pen gesture types is also a possible future work.

Another possible direction is to extend the framework to motion-sensing input, which can support gesture input with greater variations. Extending the framework to other types of devices is also a possible direction [25].

## Appendix A

# A Survey on Information Categories

The survey that we conducted regarding typical inquiries from users who are trying to navigate around Beijing is shown below:

---

Please go through the scenario and the set of interactions corresponding to it. After that, please help answer the questions below.

### Scenario

*Cindy arrives at the Beijing International Airport. She is a new visiting student from Hong Kong. She wants to go to the Training Center of Beijing University of Aeronautics and Astronautics (BUAA or Beihang) to leave her luggage. She has an lunch appointment with her mentor in Microsoft Research Asia (MSRA) in the morning and needs to visit a professor in Tsinghua University in the afternoon. She also plans to have dinner with her friends in Lotus Lane. She took out her Pocket PC, which can access information about the Haidian District and Xicheng District of Beijing as well as some traffic information updates.*

(continue...)

---

---

Interactions

- System 1: Welcome to travel enquiry system. How can I help you?
- Cindy 1: I am in Beijing International Airport. How can I go to BeiHang?
- System 2: We suggest that you take the airport bus or a taxi to the Beijing University of Aeronautics and Astronautics.
- Cindy 2: Taxi is better. Please show me the fastest way to walk to MSRA then.
- System 3: I'm sorry that I do not know MSRA. May I know the location?
- Cindy 3: hmm, Sigma Center please.
- System 4: Here is the suggested path on foot.
- Cindy 4: Afterwards, how can I go here <circle=FIT Building of Tsinghua University>?
- System 5: Do you wish to get there by walking, subway, bus or taxi?
- Cindy 5: Subway please.
- System 6: Here is the information. Please get off at Wudaokou station and then walk westwards about ten minutes.
- Cindy 6: Can I walk to here <point=Peking University>?
- System 7: It takes about thirty minutes.
- Cindy 7: Oh! No. Too far for me! How about Lotus Lane?
- System 8: Do you want to go there by subway?
- Cindy 8: Sure.
- System 9: Here is the information. Please get off at Jishuitan station, take exit C than turn left walk to the south about fifteen minutes.
- Cindy 9: Thank you very much. Good-bye.
- System 10: Thank you for using the system. Good-bye.

---

(continue...)

---

Questions

1. Who are the users?
  2. What are their needs?
  3. What kind of inputs can be supported by the system?  
(e.g. speech, pen gesture, facial expression, body gesture, etc.)
  4. Which language(s) can be supported by the system?
  5. Please write down *three* kinds of information that you expect the system can provide.
-

## Appendix B

# User Tasks for Data Collection

Tables below show the 32 tasks listed in the instruction of data collection. The subjects follow the task given and use either speech and/or pen modalities to indicate the locations requested. In the tables, “location requested” is the location indicated in the “task” and the `LOC_TYPE` and subtype are indicated by the icons on the map.

Information category: BUS INFORMATION
Task: 查詢在崇文門東大街一百米內所有行經的公交車路線。 <i>Find out the bus routes that pass through the area of 100m from the Chongwenmen East Street.</i> Location requested: 崇文門東大街 <i>Chongwenmen East Street</i> LOC_TYPE: TRANSPORTATION subtype: <i>street</i>
Task: 查詢行經建國門內大街的所有公交車路線。 <i>The bus routes that pass through the Jianguomen Inner Street.</i> Location requested: 建國門內大街 <i>Jianguomen Inner Street</i> LOC_TYPE: TRANSPORTATION subtype: <i>street</i>

Information category: CHOICE OF VECHICLE
Task: 告知系統你希望搭公交車，不是乘地鐵。 <i>Inform the system that you want to take bus instead of railway.</i> Location requested: nil

Information category: MAP COMMANDS
Task: 將畫面推近到某一個位置。 <i>Zoom in to a specific point.</i> Location requested: nil
Task: 將畫面推近到某一個小範圍。 <i>Zoom in to a specific area.</i> Location requested: nil
Task: 將畫面拉遠，顯示更大的區域。 <i>Zoom out to show larger area of the map.</i> Location requested: nil
Task: 顯示地圖更西邊的部分。 <i>Show the west side of the map.</i> Location requested: nil



Information category: OPENING HOURS
Task: 查詢雙秀公園的開放時間。 <i>Inquire about the opening hours of the Shuangxiu Park.</i> Location requested: 雙秀公園 <i>Shuangxiu Park</i> LOC_TYPE: LEISURE FACILITIES subtype: <i>parks</i>
Task: 查詢新東安市場、東方廣場及賽特購物中心的營運時間。 <i>Inquire about the operation hours of the Xindong'an Plaza, the Oriental Plaza and the Scitech Plaza.</i> Locations requested: 新東安市場 <i>Xindong'an Plaza</i> 東方廣場 <i>Oriental Plaza</i> 賽特購物中心 <i>Scitech Park</i> LOC_TYPE: MAJOR BUILDINGS subtype: <i>shopping center</i>

Information category: RAILWAY INFORMATION
Task: 查詢東方廣場四百米範圍內有多少個地鐵站。 <i>Find out the number of railway stations within the area of 400m from the Oriental Plaza.</i> Location requested: 東方廣場 <i>Oriental Plaza</i> LOC_TYPE: MAJOR BUILDINGS subtype: <i>shopping center</i>
Task: 查詢國際飯店附近五百米範圍內所有地鐵站的名稱。 <i>Find out the name of the stations within the area of 500m from the International Hotel.</i> Location requested: 國際飯店 <i>International Hotel</i> LOC_TYPE: MAJOR BUILDINGS subtype: <i>hotel</i>
Task: 告知系統你的所在位置，並查詢最近的地鐵站名稱。 <i>Inform the system on your existing location and ask the name of the nearest station.</i> Location requested: 所在位置 <i>existing location</i> LOC_TYPE: nil subtype: nil

Information category: ROUTE FINDING

Task: 告知系統你正在中國建築文化中心，查詢從那裡到紫竹院公園，可以選擇的交通路線。

*Inform the system that you are now at the China Architectural Culture Center. Inquire about the route from the China Architectural Culture Center to the Purple Bamboo Park.*

Location requested: 中國建築文化中心 *China Architectural Culture Center*

LOC\_TYPE: LEISURE FACILITIES

subtype: *museum*

Location requested: 紫竹院公園 *Purple Bamboo Park*

LOC\_TYPE: LEISURE FACILITIES

subtype: *parks*

Task: 告知系統你正在北京郵電大學，查詢從北京郵電大學依次到北京航空航天大學、中國地質大學、北京科技大學及北京醫科大學，可以選擇的交通路線。

*Inform the system that you are now at the Beijing University of Posts and Telecommunications. Inquire about the route from the Beijing University of Posts and Telecommunications to the Beihang University, the China University of Geosciences, the University of Science and Technology Beijing, the Beijing Medical University in order.*

Locations requested: 北京郵電大學

*Beijing University of Posts and Telecommunications*

北京航空航天大學 *Beihang University*

中國地質大學 *China University of Geosciences*

北京科技大學

*University of Science and Technology Beijing*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

*Remark: A task may contain up to six locations.*

Task: 告知系列你正在北京郵電大學，查詢從那裡到中國地質大學、北京科技大學、北京航空航天大學及北京醫科大學，可以選擇的交通路線。

*Inform the system that you are now at the Beijing University of Posts and Telecommunications. Inquire about the route from the Beijing University of Posts and Telecommunications to the China University of Geosciences, the University of Science and Technology Beijing, the Beihang University, the Beijing Medical University.*

Locations requested: 北京郵電大學

*Beijing University of Posts and Telecommunications*

中國地質大學 *China University of Geosciences*

北京科技大學

*University of Science and Technology Beijing*

北京航空航天大學 *Beihang University*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

*Remark: A task may contain up to six locations.*

Task: 查詢從你身處的地點（即地圖上的x）到中國人民大學，最快捷的交通路線。

*Inquire about the fastest route from your existing location to the Renmin University of China.*

Location requested: 身處的地點 *existing location*

LOC\_TYPE: nil

subtype: nil

Location requested: 中國人民大學 *Renmin University of China*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

Task: 查詢從你身處的地點（即地圖上的x）依次走到北京航空航天大學、中國地質大學、北京科技大學及北京醫科大學，最快捷的交通路線。

*Inquire about the fastest route from your existing location to the Beihang University, the China University of Geosciences, the University of Science and Technology Beijing, the Beijing Medical University in order.*

Location requested: 身處的地點 *Existing location*

LOC\_TYPE: nil

subtype: nil

Locations requested: 北京航空航天大學 *Beihang University*

中國地質大學 *China University of Geosciences*

北京科技大學

*University of Science and Technology Beijing*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

Task: 查詢從你身處的地點（即地圖上的x）到中國地質大學、北京科技大學、北京航空航天大學及北京醫科大學，最快捷的交通路線。

*Inquire about the fastest route from your existing location to the China University of Geosciences, the University of Science and Technology Beijing, the Beihang University, the Beijing Medical University.*

Location requested: 身處的地點 *Existing location*

LOC\_TYPE: nil

subtype: nil

Locations requested: 中國地質大學 *China University of Geosciences*

北京科技大學

*University of Science and Technology Beijing*

北京航空航天大學 *Beihang University*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

## Task:

請系統建議一條經過故宮博物館、人民大會堂及中國革命歷史博物館的路線。

*Inquire about the route to walk through the Palace Museum, the Great Hall of the People and the Military Museum of Chinese People's Revolution.*

Location requested: 故宮博物館 *the of Chinese People's Revolution  
Military Museum*

LOC\_TYPE: LEISURE FACILITIES

subtype: *museum*

Location requested: 人民大會堂 *the Palace Museum*

LOC\_TYPE: PUBLIC FACILITIES AND

SERVICES

subtype: *heritage*

LOC\_TYPE: LEISURE FACILITIES

subtype: *museum*

Location requested:

中國革命歷史博物館

*the Great Hall of the People*

LOC\_TYPE: POLITICAL FEATURES

subtype: *district office*

LOC\_TYPE: LEISURE FACILITIES

subtype: *theater*



*Remarks: A location may classify into multiple location type and subtype; a location type and subtype include multiple locations.*

Task: 查詢如何由金朗大酒店到最近的地鐵站。

*Inquire about the route from the Jinlang Hotel to the nearest railway station.*

Location requested: 金朗大酒店 *Jinlang Hotel*

LOC\_TYPE: MAJOR BUILDINGS

subtype: *hotel*

Task: 要求系統建議一條需時最短的路線。

*Inform the system that you need a route which takes the shortest time.*

Location requested: nil

Information category: TIME CONSTRAINT

Task: 告知系統你要在二十分鐘內到達國際飯店。

*Inform the system that you have to arrive at the International Hotel in 20 mins.*

Location requested: 國際飯店 *International Hotel*

LOC\_TYPE: MAJOR BUILDINGS

subtype: *hotel*

Information category: TRANSPORTATION COSTS

Task: 查詢從王府井坐地鐵到建國門需要多少錢。

*Inquire about the transportation cost with railway from the Wangfujing station to the Jianhuomen station.*

Locations requested: 王府井 *Wangfujing station*

建國門 *Jianhuomen station*

LOC\_TYPE: TRANSPORTATION

subtype: *railway station*

Information category: TRAVEL TIME

Task. 查詢從你身處的地點（即地圖上的x）到中國人民大學東邊二百米，需要多少時間。

*Inquire about the travel time from your existing location to 200m east from the Renmin University of China.*

Location requested: 身處的地點 *existing location*

LOC\_TYPE: nil

subtype: nil

Location requested: 中國人民大學 *Renmin University of China*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

Task: 查詢從你身處的地點（即地圖上的x）依次走到北京航空航天大學、中國地質大學、北京科技大學及北京醫科大學，一共需要多長時間。

*Inquire about the travel time from your existing location to the Beihang University, the China University of Geosciences, the University of Science and Technology Beijing, the Beijing Medical University in order.*

Location requested: 身處的地點 *existing location*

LOC\_TYPE: nil

subtype: nil

Locations requested: 北京航空航天大學 *Beihang University*

中國地質大學 *China University of Geosciences*

北京科技大學

*University of Science and Technology Beijing*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*



Task: 查詢從你身處的地點（即地圖上的x）走路到中國地質大學、北京科技大學、北京航空航天大學及北京醫科大學，一共需要多長時間。

*Inquire about the travel time from your existing location to the China University of Geosciences, the University of Science and Technology Beijing, the Beihang University and the Beijing Medical University.*

Location requested: 身處的地點 *Existing location*

LOC\_TYPE: nil

subtype: nil

Locations requested: 中國地質大學 *China University of Geosciences*  
北京科技大學 *University of Science and Technology*

*Beijing*

北京航空航天大學

*Beihang University*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

Task: 告知系統你在中國建築文化中心，查詢從那裡到紫竹院公園，需要多少時間。

*Tell the system that you are now at the China Architectural Culture Center. Inquire about the travel time from the China Architectural Culture Center to the Purple Bamboo Park.*

Location requested: 中國建築文化中心 *China Architectural Culture Center*

LOC\_TYPE: LEISURE FACILITIES

subtype: *museum*

Location requested: 紫竹院公園 *Purple Bamboo Park*

LOC\_TYPE: LEISURE FACILITIES

subtype: *parks*

Task. 查詢從北京郵電大學依次走路到北京航空航天大學、中國地質大學、北京科技大學及北京醫科大學，一共需要多長時間。

*Inquire about the travel time from the Beijing University of Posts and Telecommunications to the Beihang University, the China University of Geosciences, the University of Science and Technology Beijing and the Beijing Medical University in order.*

Locations requested: 北京郵電大學

*Beijing University of Posts and Telecommunications*

北京航空航天大學 *Beihang University*

中國地質大學 *China University of Geosciences*

北京科技大學

*University of Science and Technology Beijing*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE: SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

Task: 查詢從北京郵電大學到中國地質大學、北京科技大學、北京航空航天大學及北京醫科大學，一共需要多長時間。

*Inquire about the travel time from the Beijing University of Posts and Telecommunications to the China University of Geosciences, the University of Science and Technology Beijing, the Beihang University and the Beijing Medical University.*

Locations requested: 北京郵電大學

*Beijing University of Posts and Telecommunications*

中國地質大學 *China University of Geosciences*

北京科技大學 *University of Science and Technology*

*Beijing*

北京航空航天大學

*Beihang University*

北京醫科大學 *Beijing Medical University*

LOC\_TYPE SCHOOL AND PUBLIC LIBRARIES

subtype: *university*

Task: 查詢由國際飯店到最近的地鐵站的所需時間。

*Inquire about the travel time from the International Hotel to the nearest railway station.*

Location requested: 國際飯店 *International Hotel*

LOC\_TYPE: MAJOR BUILDINGS

subtype: *hotel*

Task: 查詢需要多長時間才可遊覽王府井大街。

*Inquire about the time to ramble around the Wangfujing.*

Location requested: 王府井大街 *International Hotel*

LOC\_TYPE: LAND AND WATER

subtype: *occupied land*

Task: 查詢需要多長時間可走以完王府井大街。

*Inquire the time to walk through the Wangfujing Avenue.*

Location requested: 王府井大街 *International Hotel*

LOC\_TYPE: LEISURE FACILITIES

subtype: *scenic spot*

LOC\_TYPE: TRANSPORTATION

subtype: *street*

*Remarks: A location can be classified into multiple location type and subtype and vice versa.*

## Appendix C

# An Instruction Provided by a Subject

Figures below is an example of the instruction provided by one of the subjects. The subject typed the speech part of the multimodal inquiries and marked the pen gestures using a pencil. During the data collection, he revised the speech part of multimodal inquiries (e.g. inquiries 49, 50 and 51). Inquiries without pencil marking, including inquiries 64, 65 and 66, are unimodal (speech-only) inputs.

請參看每一條查詢的背景資料，包括目的、任務、限制及提供的地圖，然後設計一句合適的多模態 (multi-modal) 查詢，包括用字、輸入媒介 (語音及筆觸輸入)、輸入方法 (點、圈及線) 及時間等，請記錄你所設計的查詢及原因，並實現於掌上電腦 (PPC) 上，謝謝!

直指詞 (deictic term) 的例子：這裡，那裡，這些，那些等等。

特性形容詞 (epithetic term) 的例子：清華大學→大學，天壇公園→公園

目的：資料查詢 (其他資料)	限制：用直指詞 (deictic term)	限制：用特性形容詞 (epithetic term)	限制：不用任何指示詞 (referential expression)
任務：查詢雙秀公園的開放時間。 地圖：map-b1-2.JPG	查詢：(1) 這兒的開放時間 原因：	查詢：(2) 這個公園的開放時間 原因：	查詢：(3) 雙秀公園開放時間 原因：
任務：查詢新東安市場、東方廣場及賽特購物中心的營運時間。 地圖：map-c3-2.JPG	查詢：(4) 這幾處的營業時間 原因：	查詢：(5) 這幾處商場的營業時間 原因：	查詢：(6) 列出新東安、東方廣場還有賽特的營業時間 原因：

目的：資料查詢（時間）	限制：用直指詞（deictic term）	限制：用特性形容詞（epithetic term）	限制：不能用任何指示詞（referential expression）
任務：查詢從你身處的地點（即地圖上的 x）到中國人民大學東邊二百米，需要多少時間。 地圖：map-b1-1A.JPG	查詢：(7) 這兒到那兒要多長時間	查詢：(8) 這兒到這個大學要多少時間	查詢：(9) 從所在處到人大北邊 200 米要多長時間
任務：查詢從你身處的地點（即地圖上的 x）依次走到北京航空航大、中國地質大學、北京科技大學及北京醫科大學，一共需要多長時間。 地圖：map-b1-1A.JPG	查詢：(10) 從這兒依次經過這兒，這兒，這兒，到這兒要多長時間	查詢：(11) 從這兒依次經過這個大學，這個大學，到這個大學要多長時間	查詢：(12) 從所在處依次經過北航，地大，科大到北醫要多長時間
任務：查詢從你身處的地點（即地圖上的 x）走路到中國地質大學、北京科技大學、北京航空航大及北京醫科大學，一共需要多長時間。 地圖：map-b1-2A.JPG	查詢：(13) 從這兒到這兒，這兒，這兒，這兒一共要多長時間	查詢：(14) 從這兒到這個大學，這個大學，到這個大學一共要多長時間	查詢：(15) 從所在處到北航，地大，科大，北醫一共要多長時間

目的：資料查詢（時間）	限制：用直指詞（deictic term）	限制：用特性形容詞（epithetic term）	限制：不能用任何指示詞（referential expression）
任務：告知系統你在中國建築文化中心，查詢從那裡到紫竹院公園，需要多少時間。 地圖：map-b2-1.JPG	查詢：(16)我在中國建築文化中心，到這兒要多長時間？	查詢：(17)我在中國建築文化中心，到這個公園要多長時間？	查詢：(18)我在中國建築文化中心，到紫竹院公園要多長時間？
任務：查詢從地點 北京郵電大學 悠遊走路到 北京航空航大、中國地質大學、北京科技大學及北京醫科大學，一共需要多長時間。 地圖：map-b1-2.JPG	查詢：(19) 從這兒依次經過這兒，這兒，這兒，到這兒要多長時間？	查詢：(20) 從這個大學依次經過這個大學，這個大學，到這個大學要多長時間？	查詢：(21) 從北郵依次經過北航，地大，科大，到北醫要多長時間？
任務：查詢從北京郵電大學到中國地質大學、北京科技大學、北京航空航大及北京醫科大學，一共需要多長時間。 地圖：map-b1-2.JPG	查詢：(22) 從這兒到這兒，這兒，這兒，這兒一共要多長時間？	查詢：(23) 從這個大學到這個大學，這個大學，到這個大學一共要多長時間？	查詢：(24) 從北郵到北航，地大，科大，北醫一共要多長時間？

目的：資料查詢(交通路線選擇)	限制：用直指詞 (deictic term)	限制：用特性形容詞 (epithetic term)	限制：不能用任何指示詞 (referential expression)
任務：告知系統你正在中國建築文化中心，查詢從那裡到紫竹院公園，可以選擇的交通路線。 地圖：map-b2-1.JPG	查詢：(25)我在中國建築文化中心，到这儿都可以怎么走？	查詢：(26)我在中國建築文化中心，到這個公園都可以怎么走？	查詢：(27)我在中國建築文化中心，到紫竹院公園都可以怎么走？
任務：告知系統你正在北京郵電大學，查詢從北京郵電大學依次到北京航空航天大学、中國地質大學、北京科技大學及北京醫科大學，可以選擇的交通路線。 地圖：map-b1-2.JPG	查詢：(28)从这儿依次经过这儿，这儿，这儿到这儿都可以怎么走？	查詢：(29)从这个大学依次经过这个大学，这个大学，到这个大学都可以怎么走？	查詢：(30)从北邮依次经过北航，地大，科大，到北医都可以怎么走？
任務：告知系統你正在北京郵電大學，查詢從那裡到中國地質大學、北京科技大學、北京航空航天大学及北京醫科大學，可以選擇的交通路線。 地圖：map-b1-2.JPG	查詢：(31)我在这儿，从这儿到这儿，这儿，这儿，这儿，怎么走	查詢：(32)我在这个大学，从这个大学到这个大学，这个大学，这个大学，怎么走	查詢：(33)我在北邮，从北邮到地大，科大，北航，北医怎么走



目的：資料查詢 (路線)	限制：用直指詞 (deictic term)	限制：用特性形容詞 (epithetic term)	限制：不能用任何指示詞 (referential expression)
任務：查詢從你身處的地點 (即地圖上的 x) 到中國人民大學，最快捷的交通路線。 地圖：map-b1-1A.JPG	查詢：(34) 從所在地怎樣最快到這兒？	查詢：(35) 從所在地怎樣最快到這個大學	查詢：(36) 從所在地怎樣最快到人大？
任務：查詢從你身處的地點 (即地圖上的 x) 依次走到北京航空航大、中國地質大學、北京科技大學及北京醫科大學，最快捷的交通路線。 地圖：map-b1-1A.JPG	查詢：(37) 從這兒依次經過這兒，這兒，這兒到這兒的最快路線	查詢：(38) 從這個大學依次經過這個大學，這個大學，到這個大學的最快路線	查詢：(39) 從北郵依次經過北航，地大，科大，到北醫的最快路線
任務：查詢從你身處的地點 (即地圖上的 x) 到中國地質大學、北京科技大學、北京航空航大、北京醫科大學，最快捷的交通路線。 地圖：map-b1-2A.JPG	查詢：(40) 從這兒到這兒，這兒，這兒，這兒，這兒最快怎么走	查詢：(41) 從這個大學到這個大學，這個大學，到這個大學，最快怎么走	查詢：(42) 從這兒到地大，科大，北航，北醫，最快怎么走

目的：資料查詢（其他資料）	限制：用直指詞（deictic term）	限制：用特性形容詞（epithetic term）	限制：不能用任何指示詞（referential expression）
任務：查詢東方廣場四百米範圍內有多少個地鐵站。 地圖：map-c3-2.JPG	查詢：(43) 這兒四百米內有多少個地鐵站？	查詢：(44) 這個廣場四百米內有多少個地鐵站？	查詢：(45) 東方廣場四百米內有多少個地鐵站？
任務：查詢從王府井坐地鐵到建國門需要多少錢。 地圖：map-c3-2.JPG	查詢：(46) 從這兒到這兒地鐵多少錢？	查詢：(47) 從這一站到這一站地鐵多少錢？	查詢：(48) 從王府井到建國門多少錢？
任務：查詢國際飯店附近五百米範圍內所有地鐵站的名稱。 地圖：map-c3-2.JPG	查詢：(49) 這兒五百米內有多少個地鐵站？ 列出這兒	查詢：(50) 這兒這個飯店五百米內有多少個地鐵站？	查詢：(51) 所有國際飯店五百米內有多少個地鐵站？

<p>目的：資料查詢（時間）</p> <p>任務：查詢由國際飯店到最近的地鐵站的所需時間。</p> <p>地圖：map-c3-2.JPG</p>	<p>限制：包括筆觸輸入（pen input）</p> <p>查詢：(52) 这儿到地铁站要多长时间？</p>	<p>目的：資料查詢（路線）</p> <p>任務：查詢如何由金朗大酒店到最近的地鐵站。</p> <p>地圖：map-c3-2.JPG</p>	<p>限制：包括筆觸輸入（pen input）</p> <p>查詢：(53) 这儿到地铁站怎么走</p>
<p>任務：查詢需要多長時間才可遊覽王府井大街。</p> <p>地圖：map-c3-2.JPG</p>	<p>查詢：(54)</p> <p><del>王府井大街</del> 玩儿一圈要多久？</p> <p>这条街</p>	<p>任務：查詢在崇文門東大街一百米內所有行經的公車路線。</p> <p>地圖：map-c3-2.JPG</p>	<p>查詢：(55)</p> <p>列出所有经过这儿的公共汽车</p>
<p>任務：查詢需要多長時間可走完王府井大街。</p> <p>地圖：map-c3-2.JPG</p>	<p>查詢：(56)</p> <p><del>王府井大街</del> 从这头到那头要走多久？</p> <p>这条街</p>	<p>任務：查詢行經建國門內大街的所有公車路線。</p> <p>地圖：map-c3-2.JPG</p>	<p>查詢：(57)</p> <p>列出所有经过建國大街的公共汽车</p>

目的：資料查詢 (其他資料)	限制：包括筆觸輸入 (pen input)	目的：控制	限制：包括筆觸輸入 (pen input)
<p>任務：告知系統你的所在位置，並查詢最近的地鐵站名稱。 地圖：map-c3-2.JPG</p>	<p>查詢：(58) 我在这儿，地铁怎么走？</p>	<p>任務：將畫面推近到某一個位置。 地圖：map-c3-2.JPG</p>	<p>查詢：(59) 推进到这儿</p>
<p>任務：告知系統你要在二十分鐘內到達天壇飯店。 地圖：map-c3-2.JPG</p>	<p>查詢：(60) 怎么样在二十分鐘內到達天壇飯店？</p>	<p>任務：將畫面推近到某一個小範圍。 地圖：map-c3-2.JPG</p>	<p>查詢：(61) 推进到这儿</p>
<p>任務：請系統建議一條經過故宮博物館、人民大會堂及中國革命歷史博物館的路線。 地圖：map-c3-1.JPG</p>	<p>查詢：(62) 建議一條經過故宮博物館、人民大會堂及中國革命歷史博物館的路線。</p>	<p>任務：將畫面拉遠，顯示更大的區域。 地圖：map-c3-2.JPG</p>	<p>查詢：(63) 拉远</p>

<p>目的：錄音</p> <p>任務：告知系統你希望搭公共交通車，不是乘地鐵。</p> <p>地圖：map-c3-2.JPG</p>	<p>限制：沒有</p> <p>查詢：(64)我想坐車</p>
<p>任務：要求系統建議一條需時最短的路線。</p> <p>地圖：map-c3-2.JPG</p>	<p>查詢：(65)找一條最快的路線</p>
<p>任務：顯示地圖東西邊的部分。</p> <p>地圖：map-c3-2.JPG</p>	<p>查詢：(66)向西</p>



map-b2-1.JPG



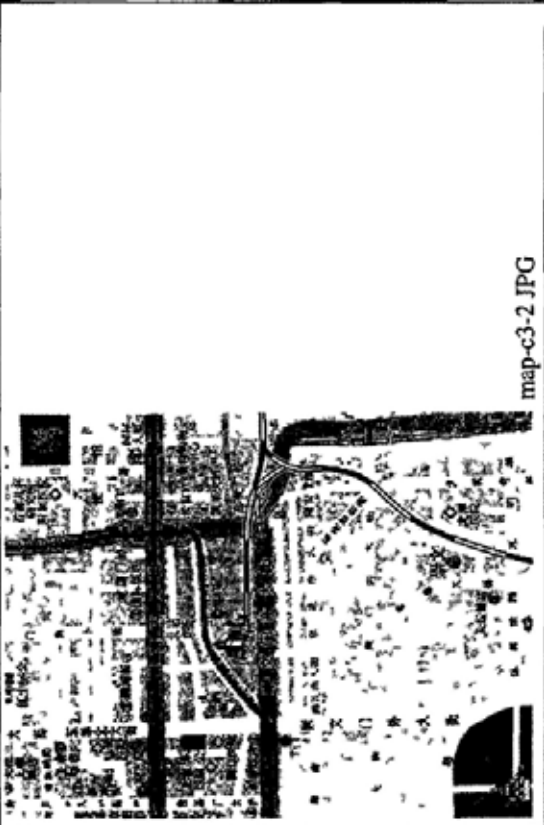
map-b1-1A.JPG



map-b1-2A.JPG



map-b1-2.JPG



map-c3-2.JPG



map-c3-1.JPG

## Appendix D

# An Illustrative Example on the Normalized Cost $C_S(S_r, N)$ for the Recognized SLR $S_r$

Table D.1 shows an illustrative example on the scoring of the recognized spoken locative references (SLRs). In this example, the first SLR has been transcribed as “這兒” *here* for 100 times across  $N$ -best speech recognition hypotheses ( $N = 100$ ). Therefore, its cost is:

$$C_S(S_r, N) = 1 - \frac{n_{S_r, N}}{N} = 1 - \frac{100}{100} = 0$$

The second SLR has been transcribed as “這幾個地方” *these places* or “這裡” *here* for 94 and 6 times respectively across  $N$ -best speech recognition hypotheses. Therefore, the cost for “這幾個地方” *these places* is:

$$C_S(S_r, N) = 1 - \frac{n_{S_r, N}}{N} = 1 - \frac{94}{100} = 0.06$$

and the cost for “這裡” *here* is:

$$C_S(S_r, N) = 1 - \frac{n_{S_r, N}}{N} = 1 - \frac{6}{100} = 0.94$$

<b>Reference transcription</b>
從“這兒”依次走到“這幾個地方”一共需要多久
<i>How much time will it take from “here”, to “these places” in sequence?</i>

<b>Speech recognition hypotheses</b>	
1	從“這兒”依次走到“這幾個地方”你一共需要多久
2	從“這兒”依次走到“這幾個地方”給共需要多久
3	從“這兒”依次走到“這幾個地方”給公交多久
4	從“這兒”依次走到“這幾個地方”米一共需要多久
5	從“這兒”依次走到“這幾個地方”裡一共需要多久
6	從“這兒”依次走到“這幾個地方”你提供需要多久
7	從“這兒”依次走到“這幾個地方”你移共需要多久
8	從“這兒”依次走到“這幾個地方”細一共需要多久
9	從“這兒”依次走到“這幾個地方”百一共需要多久
10	從“這兒”依次走到“這幾個地方”給鐘需要多久
11	鐘“這兒”依次走到“這幾個地方”你一共需要多久
12	東“這兒”依次走到“這幾個地方”你一共需要多久
13	從“這兒”依次走到“這幾個地方”給一共需要多久
14	從“這兒”依次走到“這幾個地方”你一共條多久
15	鐘“這兒”依次走到“這幾個地方”給共需要多久
16	東“這兒”依次走到“這幾個地方”給共需要多久
17	從“這兒”依次走到“這幾個地方”你移公交多久
18	從“這兒”依次走到“這幾個地方”米提供需要多久
19	從“這兒”依次走到“這幾個地方”裡提供需要多久
20	總共“這兒”依次走到“這幾個地方”你一共需要多久



21	從“這兒”依次走到“這幾個地方”你及共需要多久
22	總共“這兒”依次走到“這幾個地方”給共需要多久
23	從“這兒”依次走到問“這幾個地方”你一共需要多久
24	從“這兒”依次走到“這幾個地方”米移共需要多久
25	從“這兒”依次走到“這幾個地方”你一共要多久
26	從“這兒”依次走到“這幾個地方”裡移共需要多久
27	從“這兒”依次走到覽“這幾個地方”你一共需要多久
28	從“這兒”依次走到“這幾個地方”百提供需要多久
29	從“這兒”依次走到“這幾個地方”給共需一多久
30	從“這兒”依次走到“這幾個地方”細提供需要多久
31	從“這兒”依次走到問“這幾個地方”給共需要多久
32	從“這兒”依次走到完“這幾個地方”你一共需要多久
33	從“這兒”依次走到覽“這幾個地方”給共需要多久
34	鐘“這兒”依次走到“這幾個地方”給公交多久
35	東“這兒”依次走到“這幾個地方”給公交多久
36	從“這兒”依次走到“這幾個地方”點一共需要多久
37	從“這兒”依次走到完“這幾個地方”給共需要多久
38	從“這兒”依次走到“這幾個地方”細移共需要多久
39	從“這兒”依次走到“這幾個地方”轉一共需要多久
40	鐘“這兒”依次走到“這幾個地方”米一共需要多久
41	鐘“這兒”依次走到“這幾個地方”裡一共需要多久
42	從“這兒”依次走到玩“這幾個地方”你一共需要多久
43	東“這兒”依次走到“這幾個地方”米一共需要多久
44	東“這兒”依次走到“這幾個地方”裡一共需要多久

45	從“這兒”依次走到“這幾個地方”你離共需要多久
46	從“這兒”依次走到“這幾個地方”你移鐘需要多久
47	從“這兒”依次走到“這幾個地方”百移共需要多久
48	鐘“這兒”依次走到“這幾個地方”你提供需要多久
49	東“這兒”依次走到“這幾個地方”你提供需要多久
50	從“這兒”依次走到“這幾個地方”你提供條多久
51	從“這兒”依次走到“這幾個地方”你一共需要多久
52	從“這兒”依次走到“這幾個地方”你及公交多久
53	總共“這兒”依次走到“這幾個地方”給公交多久
54	從“這兒”依次走到“這幾個地方”米一共條多久
55	從“這兒”依次走到“這幾個地方”裡一共條多久
56	從“這兒”依次走到“這幾個地方”鐵一共需要多久
57	從“這兒”依次走到玩“這幾個地方”給共需要多久
58	從“這兒”依次走到“這幾個地方”米移公交多久
59	從“這兒”依次走到“這幾個地方”裡移公交多久
60	從“這兒”依次走到“這幾個地方”給提供需要多久
61	總共“這兒”依次走到“這幾個地方”米一共需要多久
62	總共“這兒”依次走到“這幾個地方”裡一共需要多久
63	從“這兒”依次走到“這幾個地方”給共需要多久
64	從“這兒”依次走到“這幾個地方”往你一共需要多久
65	從“這兒”依次走到問“這幾個地方”給公交多久
66	總共“這兒”依次走到“這幾個地方”你提供需要多久
67	從“這兒”依次遠走到“這幾個地方”你一共需要多久
68	從“這兒”依次走到最地鐵“這裡”往你一共需要多久

69	從“這兒”依次走到“這幾個地方”往給共需要多久
70	從“這兒”依次走到最近給“這裡”往你一共需要多久
71	鐘“這兒”依次走到“這幾個地方”你移共需要多久
72	從“這兒”依次走到給地鐵“這裡”往你一共需要多久
73	從“這兒”依次走到覽“這幾個地方”給公交多久
74	東“這兒”依次走到“這幾個地方”你移共需要多久
75	鐘“這兒”依次走到“這幾個地方”細一共需要多久
76	東“這兒”依次走到“這幾個地方”細一共需要多久
77	從“這兒”依次走到最地鐵“這裡”往給共需要多久
78	從“這兒”依次走到“這幾個地方”米及共需要多久
79	從“這兒”依次走到“這幾個地方”裡及共需要多久
80	從“這兒”依次走到最近給“這裡”往給共需要多久
81	從“這兒”依次走到給地鐵“這裡”往給共需要多久
82	從“這兒”依次走到“這幾個地方”給移共需要多久
83	從“這兒”依次走到問“這幾個地方”米一共需要多久
84	從“這兒”依次走到“這幾個地方”買一共需要多久
85	從“這兒”依次走到問“這幾個地方”裡一共需要多久
86	從“這兒”依次走到“這幾個地方”一共需要多久
87	從“這兒”依次走到“這幾個地方”米一共要多久
88	鐘“這兒”依次走到“這幾個地方”百一共需要多久
89	從“這兒”依次走到“這幾個地方”裡一共要多久
90	從“這兒”依次走到“這幾個地方”細一共條多久
91	從“這兒”依次還走到“這幾個地方”給共需要多久
92	東“這兒”依次走到“這幾個地方”百一共需要多久

93	乘“這兒”依次走到“這幾個地方”你一共需要多久
94	從“這兒”依次走到問“這幾個地方”你提供需要多久
95	從“這兒”依次走到“這幾個地方”會一共需要多久
96	從“這兒”依次走到覽“這幾個地方”米一共需要多久
97	從“這兒”依次走到覽“這幾個地方”裡一共需要多久
98	從“這兒”依次走到“這幾個地方”問一共需要多久
99	從“這兒”依次走到完“這幾個地方”給公交多久
100	從“這兒”依次走到“這幾個地方”細移公交多久

Table D.1: An example showing the normalized cost of each recognized SLR based on Equation 6.1 for the  $N$ -best ( $N = 100$ ) recognition hypotheses.

## Appendix E

# An Illustrative Example on the Hypothesis Rescoring Procedure

Table E.1 shows an illustrative example on the hypothesis rescoring procedure for candidates of cross-modality integration listed in Table D.1. The first column of Table E.1 is the rank of the speech recognition hypothesis (labeled as “SR rank”), the second column is the details of the hypothesis pair and the score obtained according to Equation 6.6 and the third column is the new rank of the hypothesis pair obtained after the hypothesis rescoring procedure (labeled as “HR rank”).

Reference transcription
S: 從“這兒”依次走到“這幾個地方”一共需要多久
P:     •                                     •••••
<i>How much time will it take from here, to these places in sequence?</i>

SR rank	Hypothesis Pairs and $C_{Tot}(S_R, P_Q)$	HR rank
1	<p><math>S</math> 從“這兒”依次走到“這幾個地方”你一共需要多久</p> <p><math>P</math>     •     • • • • •</p> $C_{Tot}(S_R, P_Q) = w_I C_I(S_R, P_Q) + w_P C_P(P_Q) + w_S C_S(S_R)$ $= 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 23.03$	8
2	<p><math>S</math> 從“這兒”依次走到“這幾個地方”給共需要多久</p> <p><math>P</math>     •     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 30.89$	20
3	<p><math>S</math> 從“這兒”依次走到“這幾個地方”給公交多久</p> <p><math>P</math>     •     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 60.44$	57
4	<p><math>S</math> 從“這兒”依次走到“這幾個地方”米一共需要多久</p> <p><math>P</math>     •     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 21.44$	5
5	<p><math>S</math> 從“這兒”依次走到“這幾個地方”裡一共需要多久</p> <p><math>P</math>     •     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 28.95$	17
6	<p><math>S</math> 從“這兒”依次走到“這幾個地方”你提供需要多久</p> <p><math>P</math>     •     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 47.14$	37

7	<p>S: 從“這兒”依次走到“這幾個地方”你移共需要多久</p> <p>P:    •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 47.93$	39
8	<p>S: 從“這兒”依次走到“這幾個地方”細一共需要多久</p> <p>P:    •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 28.95$	18
9	<p>S: 從“這兒”依次走到“這幾個地方”百一共需要多久</p> <p>P:    •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 21.68$	6
10	<p>S: 從“這兒”依次走到“這幾個地方”給鐘需要多久</p> <p>P:    •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 47.69$	38
11	<p>S: 鐘“這兒”依次走到“這幾個地方”你一共需要多久</p> <p>P:    •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 60.63$	58
12	<p>S: 東“這兒”依次走到“這幾個地方”你一共需要多久</p> <p>P:    •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 63.16$	60

<p>13</p>	<p>S: 從“這兒”依次走到“這幾個地方”給一共需要多久</p> <p>P:     •                         •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 24.47$	<p>11</p>
<p>14</p>	<p>S: 從“這兒”依次走到“這幾個地方”你一共條多久</p> <p>P:     •                         •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 83.92$	<p>72</p>
<p>15</p>	<p>S: 鐘“這兒”依次走到“這幾個地方”給共需要多久</p> <p>P:     •                         •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 81.28$	<p>70</p>
<p>16</p>	<p>S: 東“這兒”依次走到“這幾個地方”給共需要多久</p> <p>P:     •                         •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 84.67$	<p>73</p>
<p>17</p>	<p>S: 從“這兒”依次走到“這幾個地方”你移公交多久</p> <p>P:     •                         •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 91.64$	<p>79</p>
<p>18</p>	<p>S: 從“這兒”依次走到“這幾個地方”米提供需要多久</p> <p>P:     •                         •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 43.87$	<p>31</p>





25	<p>S: 從“這兒”依次走到“這幾個地方”你一共要多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 23.68$	10
26	<p>S: 從“這兒”依次走到“這幾個地方”裡移共需要多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 58.99$	52
27	<p>S: 從“這兒”依次走到覽“這幾個地方”你一共需要多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 41.66$	28
28	<p>S: 從“這兒”依次走到“這幾個地方”百提供需要多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 44.37$	33
29	<p>S: 從“這兒”依次走到“這幾個地方”給共需一多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 163.97$	91
30	<p>S: 從“這兒”依次走到“這幾個地方”細提供需要多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 59.24$	55

31	<p>S: 從“這兒”依次走到問“這幾個地方”給共需要多久</p> <p>P:       •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 51.07$	43
32	<p>S: 從“這兒”依次走到完“這幾個地方”你一共需要多久</p> <p>P:       •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 33.14$	22
33	<p>S: 從“這兒”依次走到覽“這幾個地方”給共需要多久</p> <p>P:       •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 54.38$	47
34	<p>S: 鐘“這兒”依次走到“這幾個地方”給公交多久</p> <p>P:       •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 177.1207$	92
35	<p>S: 東“這兒”依次走到“這幾個地方”給公交多久</p> <p>P:       •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 185.35$	93
36	<p>S: 從“這兒”依次走到“這幾個地方”點一共需要多久</p> <p>P:       •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 21.03$	4

37	<p>S: 從“這兒”依次走到完“這幾個地方”給共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 43.27$	30
38	<p>S: 從“這兒”依次走到“這幾個地方”細移共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 58.99$	53
39	<p>S: 從“這兒”依次走到“這幾個地方”轉一共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 27.01$	12
40	<p>S: 鐘“這兒”依次走到“這幾個地方”米一共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 56.414$	49
41	<p>S: 鐘“這兒”依次走到“這幾個地方”裡一共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 76.19$	65
42	<p>S: 從“這兒”依次走到玩“這幾個地方”你一共需要多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 39.15$	26

43	<p>S 東 “這兒” 依次 走到 “這幾個地方” 米 一共 需要 多久</p> <p><math>P \quad \bullet \quad \dots \bullet</math></p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 58.77$	51
44	<p>S 東 “這兒” 依次 走到 “這幾個地方” 裡 一共 需要 多久</p> <p><math>P \quad \bullet \quad \dots \bullet</math></p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 79.37$	69
45	<p>S 從 “這兒” 依次 走到 “這幾個地方” 你 離 共需 要 多久</p> <p><math>P \quad \bullet \quad \dots \bullet</math></p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 44.35$	32
46	<p>S 從 “這兒” 依次 走到 “這幾個地方” 你 移 鐘 需 要 多 久</p> <p><math>P \quad \bullet \quad \dots \bullet</math></p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 71.14$	63
47	<p>S 從 “這兒” 依次 走到 “這幾個地方” 百 移 共需 要 多 久</p> <p><math>P \quad \bullet \quad \dots \bullet</math></p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 45.36$	36
48	<p>S 鐘 “這兒” 依次 走到 “這幾個地方” 你 提 供 需 要 多 久</p> <p><math>P \quad \bullet \quad \dots \bullet</math></p> $C_{Tot}(S_R, P_Q) = 0.5 \frac{0}{7} + 0.35 \frac{(0+0+0+0+0)}{5} + 0.15 \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 124.07$	88

49	<p>S: 東 “這兒” 依次 走到 “這幾個地方” 你 提供 需要 多久</p> <p>P:     •                                     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 129.25</math></p>	89
50	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 你 提供 條 多久</p> <p>P:     •                                     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 90.30</math></p>	77
51	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 你 一共 需要 多久</p> <p>P:     •                                     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 23.04</math></p>	9
52	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 你 及 公交 多久</p> <p>P:     •                                     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 86.208</math></p>	76
53	<p>S: 總共 “這兒” 依次 走到 “這幾個地方” 給 公交 多久</p> <p>P:     •                                     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 199.72</math></p>	94
54	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 米 一共 條 多久</p> <p>P:     •                                     • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ <p><math>PP_{MM} = 78.09</math></p>	67

55	<p>S: 從“這兒”依次走到“這幾個地方”裡一共條多久</p> <p>P: • ••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 105.4648$	83
56	<p>S: 從“這兒”依次走到“這幾個地方”鐵一共需要多久</p> <p>P: • ••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 27.04$	14
57	<p>S: 從“這兒”依次走到玩“這幾個地方”給共需要多久</p> <p>P: • ••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 51.10$	44
58	<p>S: 從“這兒”依次走到“這幾個地方”米移公交多久</p> <p>P: • ••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 85.27$	75
59	<p>S: 從“這兒”依次走到“這幾個地方”裡移公交多久</p> <p>P: • ••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 115.16$	84
60	<p>S: 從“這兒”依次走到“這幾個地方”給提供需要多久</p> <p>P: • ••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 50.08$	41

61	<p>S: 總共 “這兒” 依次 走到 “這幾個地方” 米 一共 需要 多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 62.85$	59
62	<p>S: 總共 “這兒” 依次 走到 “這幾個地方” 裡 一共 需要 多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 84.88$	74
63	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 給 共需 要 多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 30.89$	21
64	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 往 你 一共 需要 多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 39.15$	27
65	<p>S: 從 “這兒” 依次 走到 問 “這幾個地方” 給 公交 多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 98.27$	80
66	<p>S: 總共 “這兒” 依次 走到 “這幾個地方” 你 提供 需要 多久</p> <p>P:     •   • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 138.23$	90



67	<p>S. 從“這兒”依次遠走到“這幾個地方”你一共需要多久</p> <p><i>P</i> . . . . .</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 19.73$	3
68	<p>S: 從“這兒”依次走到最地鐵“這裡”往你一共需要多久</p> <p><i>P</i>: . . . . .</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 79.69$	95
69	<p>S. 從“這兒”依次走到“這幾個地方”往給共需要多久</p> <p><i>P</i>. . . . .</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 51.11$	45
70	<p>S: 從“這兒”依次走到最近給“這裡”往你一共需要多久</p> <p><i>P</i>. . . . .</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 102.15$	98
71	<p>S. 鐘“這兒”依次走到“這幾個地方”你移共需要多久</p> <p><i>P</i>. . . . .</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 115.51$	86
72	<p>S: 從“這兒”依次走到給地鐵“這裡”往你一共需要多久</p> <p><i>P</i>. . . . .</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 84.15$	96

73	<p>S: 從“這兒”依次走到覽“這幾個地方”給 公交 多久</p> <p>P:     •                                   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 105.31$	81
74	<p>S: 東“這兒”依次走到“這幾個地方”你 移 共需要 多久</p> <p>P:     •                                   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 119.88$	87
75	<p>S: 鐘“這兒”依次走到“這幾個地方”細 一共 需要 多久</p> <p>P:     •                                   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 76.19$	66
76	<p>S: 東“這兒”依次走到“這幾個地方”細 一共 需要 多久</p> <p>P:     •                                   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 79.37$	68
77	<p>S: 從“這兒”依次走到最地鐵“這裡”往 給 共需要 多久</p> <p>P:     •                                   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 99.8442$	97
78	<p>S: 從“這兒”依次走到“這幾個地方”米 及 共需要 多久</p> <p>P:     •                                   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 42.46$	29

79	<p>S: 從“這兒”依次走到“這幾個地方”裡及共需要多久</p> <p>P:     •                                 •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 55.80$	48
80	<p>S: 從“這兒”依次走到最近給“這裡”往給共需要多久</p> <p>P:     •                                 •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 127.99$	100
81	<p>S: 從“這兒”依次走到給地鐵“這裡”往給共需要多久</p> <p>P:     •                                 •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.94)}{2} = 0.0705$ $PP_{MM} = 105.44$	99
82	<p>S: 從“這兒”依次走到“這幾個地方”給移共需要多久</p> <p>P:     •                                 •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 50.63$	42
83	<p>S: 從“這兒”依次走到問“這幾個地方”米一共需要多久</p> <p>P:     •                                 •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 35.54$	23

84	<p>S: 從“這兒”依次走到“這幾個地方”買一共需要多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 27.04$	15
85	<p>S: 從“這兒”依次走到問“這幾個地方”裡一共需要多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 48.15$	40
86	<p>S: 從“這兒”依次走到“這幾個地方”一共需要多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 18.42$	2
87	<p>S: 從“這兒”依次走到“這幾個地方”米一共要多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 22.03$	7
88	<p>S: 鐘“這兒”依次走到“這幾個地方”百一共需要多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 57.07$	50
89	<p>S: 從“這兒”依次走到“這幾個地方”裡一共要多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 29.76$	19
90	<p>S: 從“這兒”依次走到“這幾個地方”細一共條多久</p> <p>P:      •                                      • • • • •</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 105.46$	82

91	<p>S: 從 “這兒” 依次 遠走到 “這幾個地方” 給 共需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 27.33$	16
92	<p>S: 東 “這兒” 依次 走到 “這幾個地方” 百 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 59.45$	56
93	<p>S: 乘 “這兒” 依次 走到 “這幾個地方” 你 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 67.68$	62
94	<p>S: 從 “這兒” 依次 走到 問 “這幾個地方” 你 提供 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 75.02$	64
95	<p>S: 從 “這兒” 依次 走到 “這幾個地方” 會 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 27.02$	13
96	<p>S: 從 “這兒” 依次 走到 覽 “這幾個地方” 米 一共 需要 多久</p> <p>P:     •   •••••</p> $C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045$ $PP_{MM} = 39.02$	24

97	<p>S. 從“這兒”依次走到覽“這幾個地方”裡一共需要多久</p> <p>P:       •                               •••••</p> <p><math>C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045</math></p> <p><math>PP_{MM} = 51.28</math></p>	46
98	<p>S. 從“這兒”依次走到“這幾個地方”間一共需要多久</p> <p>P:       •                               •••••</p> <p><math>C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045</math></p> <p><math>PP_{MM} = 3.03</math></p>	1
99	<p>S. 從“這兒”依次走到完“這幾個地方”給公交多久</p> <p>P:       •                               •••••</p> <p><math>C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045</math></p> <p><math>PP_{MM} = 81.89</math></p>	71
100	<p>S. 從“這兒”依次走到“這幾個地方”細移公交多久</p> <p>P:       •                               •••••</p> <p><math>C_{Tot}(S_R, P_Q) = 0.5 \cdot \frac{0}{7} + 0.35 \cdot \frac{(0+0+0+0+0)}{5} + 0.15 \cdot \frac{(0+0.06)}{2} = 0.0045</math></p> <p><math>PP_{MM} = 115.16</math></p>	85

Table E.1: An example illustrating the hypothesis rescoring process of based on the  $N$ -best speech recognition hypotheses ( $N = 100$ ) listed in Table D.1. The first and the second SLRs *here*, should have the numeric feature NUM=nil, which can be aligned with any number of pen gesture. The second SLR, *these places*, should have the numeric feature NUM=plural, which can be aligned with more than one pen gestures. All the five pen gestures incur no cost because their coordinates coincide with the respect icons/labels. Each candidate for cross-modality integration is rescored and then the updated rank is shown for each candidate. The 98<sup>th</sup> hypothesis pair ranked top after rescoring.

## Appendix F

### Significance Tests

Different numbers of speech recognition hypothesis ( $N = 1$  or  $100$ ) and pen recognition hypothesis ( $M = 1$  or  $4$ ) are used in the rescoring procedure. Results are compared and the statistical significance of the differences are assessed.

### F.1 Cross-modality hypotheses rescoring of the First Best Recognized Speech Hypotheses and $M$ -Best Pen Recognition Outputs ( $M = 4$ )

We have performed significance test on cross-modality hypotheses rescoring of the first best recognized speech hypotheses and  $M$ -best pen recognition outputs ( $M = 4$ ) from the training set. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best pen recognition hypothesis and  $M$ -best pen recognition hypotheses is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f968})$  and  $r_m = (r_{m1}, r_{m2}, \dots, r_{m968})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m968} - r_{f968})$ . Table F.1 shows the procedures for the significance test on cross-modality hypotheses rescoring of the first best recognized speech hypotheses and  $M$ -best pen recognition outputs ( $M = 4$ ) from the training set.



The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.032888$  with sample deviation  $\sigma_{r_d} = 0.178435$ .

The parameter of interest is  $\mu$ , which is the mean difference between the performance for each multimodal inquiry with  $M$ -best pen recognition hypotheses and first best pen recognition hypothesis.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.032888$ ,  $\sigma_{r_d} = 0.178435$  and  $n = 968$ ,

$$z_0 = \frac{0.032888 - 0}{0.178435/\sqrt{968}} = 5.81$$

Since  $z_0 = 5.81 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $M$ -best and first best pen recognition outputs differs from 0. The experiments are performed based on first best speech recognition hypothesis and  $M$ -best pen recognition outputs in a sample of 968 multimodal inquiries from the training set.

Table F.1: A significant test on the cross-modality hypotheses rescoring of the first best recognized speech hypotheses and  $M$ -best pen recognition outputs ( $M = 4$ ) from the training set.

We have performed significance test on cross-modality hypotheses rescoring of the first best recognized speech hypotheses and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best pen recognition hypothesis and  $M$ -best pen recognition hypotheses is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f434})$  and  $r_m = (r_{m1}, r_{m2}, \dots, r_{m434})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m434} - r_{f434})$ . Table F.2 shows the procedures for the significance test on cross-modality hypotheses rescoring of the first best recognized speech hypotheses and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set.

The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.051282$  with sample deviation  $\sigma_{r_d} = 0.22083$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $M$ -best pen recognition hypotheses and first best pen recognition hypothesis.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.051282$ ,  $\sigma_{r_d} = 0.22083$  and  $n = 434$ ,

$$z_0 = \frac{0.051282 - 0}{0.22083/\sqrt{434}} = 4.84$$

Since  $z_0 = 4.84 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $M$ -best and first best pen recognition outputs differs from 0. The experiments are performed based on first best speech recognition hypothesis and  $M$ -best pen recognition outputs in a sample of 434 multimodal inquiries from the test set.

Table F.2: A significant test on the cross-modality hypotheses rescoring of the first best recognized speech hypotheses and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set.

## F.2 Cross-modality hypotheses rescoring of the $N$ -Best Speech Recognition Hypotheses ( $N = 100$ ) and First Best Pen Recognition Outputs

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the training set. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech recognition hypothesis and  $N$ -best speech recognition hypotheses is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f968})$  and  $r_m = (r_{m1}, r_{m2}, \dots, r_{m968})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m968} - r_{f968})$ . Table F.3 shows the procedures for the significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the training set.

The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.138746$  with sample deviation  $\sigma_{r_d} = 0.398388$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech recognition hypotheses and first best speech recognition hypothesis.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.138746$ ,  $\sigma_{r_d} = 0.398388$  and  $n = 968$ ,

$$z_0 = \frac{0.138746 - 0}{0.398388/\sqrt{968}} = 10.84$$

Since  $z_0 = 10.84 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $N$ -best and first best speech recognition hypotheses differs from 0. The experiments are performed based on  $N$ -best speech recognition hypotheses and first best pen recognition outputs in a sample of 968 multimodal inquiries from the training set.

Table F.3: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the training set.

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the test set. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech recognition hypothesis and  $N$ -best speech recognition hypotheses is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f434})$  and  $r_m = (r_{m1}, r_{m2}, \dots, r_{m434})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m434} - r_{f434})$ . Table F 4 shows the procedures for the significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the test set

The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.146853$  with sample deviation  $\sigma_{r_d} = 0.379831$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech recognition hypotheses and first best speech recognition hypothesis.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d} / \sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.146853$ ,  $\sigma_{r_d} = 0.379831$  and  $n = 434$ ,

$$z_0 = \frac{0.146853 - 0}{0.379831 / \sqrt{434}} = 8.05$$

Since  $z_0 = 8.05 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $N$ -best and first best speech recognition outputs differs from 0. The experiments are performed based on  $N$ -best speech recognition hypotheses and first best pen recognition outputs in a sample of 434 multimodal inquiries from the test set.

Table F.4: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and first best pen recognition outputs from the test set.

### F.3 Cross-modality hypotheses rescoring of the $N$ -Best Speech Recognition Hypotheses ( $N = 100$ ) and $M$ -Best Pen Recognition Outputs ( $M = 4$ )

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the training set. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech recognition hypotheses and first best pen recognition outputs is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f968})$ . The performance obtained for each multimodal inquiry with  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) is  $r_m = (r_{m1}, r_{m2}, \dots, r_{m968})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m968} - r_{f968})$ . Table F.5 shows the procedures for the significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the training set.



The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.168551$  with sample deviation  $\sigma_{r_d} = 0.423533$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech and  $M$ -best pen recognition hypotheses and first best speech and pen recognition hypotheses.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.168551$ ,  $\sigma_{r_d} = 0.423533$  and  $n = 968$ ,

$$z_0 = \frac{0.168551 - 0}{0.423533/\sqrt{968}} = 12.38$$

Since  $z_0 = 12.38 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $N$ -best speech and  $M$ -best pen hypotheses and first best speech and first best pen recognition outputs differs from 0. The experiments are performed based on  $N$ -best speech recognition hypotheses and  $M$ -best pen recognition outputs in a sample of 968 multimodal inquiries from the training set.

Table F.5: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the training set.

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech and first best pen recognition hypotheses is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f434})$ . The performance obtained for each multimodal inquiry with  $N$ -best speech and  $M$ -best pen recognition hypotheses is  $r_m = (r_{m1}, r_{m2}, \dots, r_{m434})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m434} - r_{f434})$ . Table F.6 shows the procedures for the significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set.

#### **F.4 Improvements in the Integration Accuracy brought about by Cross-Modality Hypotheses Rescoring in the Presence of Speech Recognition Errors**

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of speech recognition errors. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech recognition hypotheses and first best pen recognition outputs is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f260})$ . The performance obtained for each multimodal inquiry with  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) is  $r_m = (r_{m1}, r_{m2}, \dots, r_{m260})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m260} - r_{f260})$ . Table F.7 shows the procedures for the significance test on cross-modality hy-

The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.204651$  with sample deviation  $\sigma_{r_d} = 0.415298$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech and  $M$ -best pen recognition hypotheses and first best speech and first best pen recognition hypotheses.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.204651$ ,  $\sigma_{r_d} = 0.415298$  and  $n = 434$ ,

$$z_0 = \frac{0.204651 - 0}{0.415298/\sqrt{434}} = 10.27$$

Since  $z_0 = 10.27 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $N$ -best speech and  $M$ -best pen recognition hypotheses and first best speech and first best pen recognition outputs differs from 0. The experiments are performed based on  $N$ -best speech recognition hypotheses and  $M$ -best pen recognition outputs in a sample of 434 multimodal inquiries from the test set.

Table F.6: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set.

potheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of speech recognition errors.

The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.264591$  with sample deviation  $\sigma_{r_d} = 0.441976$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech and  $M$ -best pen recognition hypotheses and first best speech and pen recognition hypotheses.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.01$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d} / \sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.264591$ ,  $\sigma_{r_d} = 0.441976$  and  $n = 260$ ,

$$z_0 = \frac{0.264591 - 0}{0.441976 / \sqrt{260}} = 9.65$$

Since  $z_0 = 9.65 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.01 level of significance.

We conclude that the mean difference between the performance with  $N$ -best speech and  $M$ -best pen hypotheses and first best speech and first best pen recognition outputs differs from 0 (i.e. improvement in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from 33.6% to 56.1% in the presence of speech recognition errors). The experiments are performed based on  $N$ -best speech recognition hypotheses and  $M$ -best pen recognition outputs in a sample of 260 multimodal inquiries from the test set in the presence of speech recognition errors.

Table F.7: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of speech recognition errors.

### F.5 Improvements in the Integration Accuracy brought about by Cross-Modality Hypotheses Rescoring in the Presence of Pen Recognition Errors

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of pen recognition errors. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech recognition hypotheses and first best pen recognition outputs is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f42})$ . The performance obtained for each multimodal in-

quiry with  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) is  $r_m = (r_{m1}, r_{m2}, \dots, r_{m42})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m42} - r_{f42})$ . Table F.8 shows the procedures for the significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of pen recognition errors.

The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.243902$  with sample deviation  $\sigma_{r_d} = 0.434769$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech and  $M$ -best pen recognition hypotheses and first best speech and pen recognition hypotheses.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.001$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.005} = 2.58$  or if  $z_0 < -z_{0.005} = -2.58$ .

Since  $\bar{r}_d = 0.243902$ ,  $\sigma_{r_d} = 0.434769$  and  $n = 42$ ,

$$z_0 = \frac{0.243902 - 0}{0.434769/\sqrt{42}} = 3.64$$

Since  $z_0 = 3.64 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.001 level of significance.

We conclude that the mean difference between the performance with  $N$ -best speech and  $M$ -best pen hypotheses and first best speech and first best pen recognition outputs differs from 0 (i.e. the improvement in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant in the presence of pen recognition errors). The experiments are performed based on  $N$ -best speech recognition hypotheses and  $M$ -best pen recognition outputs in a sample of 42 multimodal inquiries from the test set in the presence of pen recognition errors.

Table F.8: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of pen recognition errors.

## F.6 Improvements in the Integration Accuracy brought about by Cross-Modality Hypotheses Rescoring in the Presence of both Speech and Pen Recognition Errors

We have performed significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of both speech and pen recognition errors. We have formulated a paired  $Z$ -test to test the significance of the experimental results. The performance obtained for each multimodal inquiry with first best speech recognition hypotheses and first best pen recognition outputs is  $r_f = (r_{f1}, r_{f2}, \dots, r_{f34})$ . The performance obtained for each multimodal inquiry with  $N$ -best speech recogni-

tion hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) is  $r_m = (r_{m1}, r_{m2}, \dots, r_{m34})$  respectively. The difference between the two results sets is  $r_d = (r_{m1} - r_{f1}, r_{m2} - r_{f2}, \dots, r_{m34} - r_{f34})$ . Table F.9 shows the procedures for the significance test on cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of both speech and pen recognition errors.



The sample mean of the performance's difference is equals to  $\bar{r}_d = 0.264706$  with sample deviation  $\sigma_{r_d} = 0.447811$ .

The parameter of interest is  $\mu$ , the mean difference between the performance for each multimodal inquiry with  $N$ -best speech and  $M$ -best pen recognition hypotheses and first best speech and pen recognition hypotheses.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\alpha = 0.001$$

The test statistic is  $z_0 = \frac{\bar{r}_d - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject  $H_0$  if  $z_0 > z_{0.05} = 2.58$  or if  $z_0 < -z_{0.05} = -2.58$ .

Since  $\bar{r}_d = 0.264706$ ,  $\sigma_{r_d} = 0.447811$  and  $n = 34$ ,

$$z_0 = \frac{0.264707 - 0}{0.447811/\sqrt{34}} = 3.45$$

Since  $z_0 = 3.45 > 2.58$ , we reject  $H_0 : \mu = 0$  at the 0.001 level of significance.

We conclude that the mean difference between the performance with  $N$ -best speech and  $M$ -best pen hypotheses and first best speech and first best pen recognition outputs differs from 0 (i.e. improvement in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant in the presence of both speech and pen recognition errors). The experiments are performed based on  $N$ -best speech recognition hypotheses and  $M$ -best pen recognition outputs in a sample of 34 multimodal inquiries from the test set in the presence of both speech and pen recognition errors.

Table F.9: A significant test on the cross-modality hypotheses rescoring of the  $N$ -best speech recognition hypotheses ( $N = 100$ ) and  $M$ -best pen recognition outputs ( $M = 4$ ) from the test set in the presence of both speech and pen recognition errors.

## Appendix G

# Abbreviations

Table G.1 includes abbreviations that occur in this thesis for quick reference.

---

ARG	Attribute Relational Graph
BUAA	Beijing University of Aeronautics and Astronautics
BUPT	Beijing University of Post and Telecommunications
CAC	Character Auto-Completion
CMI	Cross-Modality Integration
CMIP	Cross-Modal Integration Pattern
CUBRICON	CUBRC Intelligent CONversationalist
FST	Finite-State Transducer
GPS	Global Positioning System
HCWP	Human-Centric Word Processor
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IDF	Inverse Document Frequency
IR	Information Retrieval

---

---

LDA	Latent Dirichlet Allocation
LSM	Latent Semantic Modeling
MATCH	Multimodal Access To City Help
MiPAD	Multimodal Interactive Personal Assistance Device
MM	Multimodal
MSRA	Microsoft Research Asia
NLG	Natural Language Generation
NLU	Natural Language Understanding
PLSA	Probabilistic Latent Semantic Analysis
POSTech	POSTech Multimodal Dialog System
PP	Perplexity
SDS	Spoken Dialog System
SEQ	Sequential
SIM	Simultaneous
SLR	Spoken Locative Reference
SVD	Singular Value Decomposition
TF	Term Frequency
UM	Unimodal
WITAS	Wallenberg laboratory for research on Information Technology and Autonomous Systems

---

Table G.1: A list of abbreviations used in this thesis.

# Bibliography

- [1] McNeill, D.. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, USA, 1992.
- [2] Apple® iPhone  
<http://www.apple.com/iphone>
- [3] Google Nexus One  
<http://www.google.com/phone>
- [4] HP iPAQ  
<http://www.hp.com/sbso/busproducts.handhelds.html>
- [5] Bolt, R. A.. “Put-That-There”, Voice and Gesture at the Graphics Interface. In *Proceedings of 7<sup>th</sup> Annual Conference on Computer Graphics and Interactive Techniques*, pages 262–270, Seattle, Washington, USA, 14–18 July 1980. (Also published in *ACM SIGGRAPH Computer Graphics*, pages 262-270, Volume 14, Issue 3, July 1980.)
- [6] Luettn, J., G. Potamianos and C. Neti. Asynchronous Stream Modeling for Large Vocabulary Audio-Visual Speech Recognition. in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 169–172, Volume 1, Salt Lake City, USA, 7-11 May 2001.
- [7] Potamianos, G., C. Neti, J. Luettn and I. Matthews. Audio-Visual Automatic Speech Recognition: An Overview. *Issues in Visual and Audio-*

- Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson and P. Perrier (Eds.), MIT Press, 2004.
- [8] Liu, P., L. Ma and F. Soong. Prefix Tree Based Auto-Completion for Convenient Bi-modal Chinese Character Input. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4465–4468, Las Vegas, Nevada, USA, 31 March - 4 April 2008.
- [9] Cohen, P. R., M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen and J. Clow. Quickset: Multimodal Interaction for Distributed Applications. In *Proceedings of the 5<sup>th</sup> ACM International Multimedia Conference*, pages 31–40, Seattle, Washington, USA, 9-13 November 1997.
- [10] Lemon, O., A. Bracy, A. Gruenstein and S. Peters. The WITAS Multi-Modal Dialogue System I. In *Proceedings of the 7<sup>th</sup> European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1559–1562, Aalborg, Denmark, 3-7 September 2001.
- [11] WITAS  
<http://www.ida.liu.se/ext/witas/eng.html>
- [12] SmartKom  
<http://www.smartkom.org/>
- [13] Wahlster, W.. *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer Berlin Heidelberg, September 2006.
- [14] Google Maps  
<http://maps.google.com>
- [15] Oviatt, S., P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro. Designing

- the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions. *Human-Computer Interaction*, Volume 15, Issue 4, pages 263–322, 2000.
- [16] Hauptmann, A. G.. Speech and Gestures for Graphic Image Manipulation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems Conference (CHI89)*, Volume 20, pages 241–245, Texas, USA, 30 April-4 June 1989.
- [17] Oviatt, S., A. DeAngeli and K. Kuhn. Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 415–422, Atlanta, Georgia, USA, 22-27 March 1997.
- [18] Gibbon, D., I. Mertins and R. K. Moore. Audio-visual and Multimodal Speech-based Systems. *Handbook of Multimodal and Spoken Dialogue Systems - Resources, Terminology and Product Evaluation*, pages 102–203, Kluwer Academic Publishers, 2000.
- [19] Neal, J. G., C. Y. Thielman, Z. Dobes, S. M. Haller and S. C. Shapiro. Natural Language with Integrated Deictic and Graphic Gestures. *Readings in Intelligent User Interfaces*, pages 37–52, 1998.
- [20] Cheema, S. and J. J. LaViola Jr.. Towards Intelligent Motion Inferencing in Mathematical Sketching. In *Proceedings of the 14<sup>th</sup> International Conference on Intelligent User Interfaces*, pages 289–292, Hong Kong, China, 7-10 February 2010.
- [21] Johnston, M., S. Bangalore, A. Stent, G. Vasireddy and P. Ehlen. Multimodal Language Processing for Mobile Information Access. In *Proceedings of the 7<sup>th</sup> International Conference on Spoken Language Processing*

- (*ICSLP-INTERSPEECH*), pages 2237–2240, Denver, Colorado, USA, 16–20 September 2002.
- [22] Chai, J. Y., P. Hong and M. X. Zhou. A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces. In *Proceedings of the 9<sup>th</sup> International Conference on Intelligent User Interfaces (IUI)*, pages 70–77, Funchal, Portugal, 13–16 January 2004.
- [23] Wang, S.. *A Multimodal Galaxy-based Geographic System*. S. M. Thesis, MIT, 2003.
- [24] MIT Project Voyager  
<http://www.media.mit.edu/pia/voyager/>
- [25] Johnston, M. and S. Bangalore. Multimodal Applications from Mobile to Kiosk. In *Proceedings of the W3C Workshop on Multimodal Interaction*, Sophia Antipolis, France, 19–20 July 2004.
- [26] Vergo, J.. A Statistical Approach to Multimodal Natural Language Interaction. In *Proceedings of AAAI Workshop*, Technical Report WS-98-09, pages 81–85, Madison, Wisconsin, 26–27 July 1998.
- [27] Doherty, P., G. Granlund, K. Kuchcinski, E. Sandewall, K. Nordberg, E. Skarman and J. Wiklund. The WITAS Unmanned Aerial Vehicle Project. In *Proceedings of the 14<sup>th</sup> European Conference on Artificial Intelligence (ECAI)*, pages 747–755, Berlin, Germany, 20–25 August 2000.
- [28] Wang, K.. From Multimodal to Natural Interactions. In *Proceedings of the W3C Workshop on Multimodal Interaction*, Sophia Antipolis, France, 19–20 July 2004.
- [29] Miki, M., C. Miyajima, T. Nishino, N. Kitaoka and K. Takeda. An Integrative Recognition Method for Speech and Gestures. In *Proceedings of*



- the 10<sup>th</sup> International Conference on Multimodal Interfaces (ICMI)*, pages 93–96, Chania, Greece, 20-22 October 2008.
- [30] Boves, L., A. Neumann, L. Vuurpijl, L. ten Bosch, S. Rossignol, R. Engel, and N. Pflieger. Multimodal Interaction in Architectural Design Applications. In *Proceedings of UI4ALL 2004: 8th ERCIM Workshop on "User Interfaces for All"*, Vienna, Austria, 28-29 June 2004.
- [31] COMIC - CONversational Multimodal Interaction with Computers  
<http://www.hrc.ed.ac.uk/comic/>
- [32] Zeng Z., J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. S. Huang, D. Roth and S. Levinson. Bimodal HCI-related Affect Recognition. In *Proceedings of the 6<sup>th</sup> International Conference on Multimodal Interfaces (ICMI)*, pages 137–143, State College, PA, USA, 13-15 October 2004.
- [33] Nakano, Y. I. and R. Ishii. Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations. In *Proceedings of International Conference on Intelligent User Interfaces (IUI)*, pages 139–148, Hong Kong, China, 7-10 February 2010.
- [34] Hui, H., H. Meng and M. W. Mak. Adaptive Weight Estimation in Multi-Biometric Verification using Fuzzy Logic Decision Fusion. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 1, pages 501–504, Hawai'i Convention Center, Honolulu, Hawaii, USA, 15-20 April 2007.
- [35] Morency, L. P., C. Sidner, C. Lee and T. Darrell. Contextual Recognition of Head Gestures. In *Proceedings of the 7<sup>th</sup> International Conference on Multimodal Interfaces (ICMI)*, pages 18–24, Trento, Italy, 4-6 October 2005.

- [36] Ko, T., D. Demirdjian and T. Darrell. Untethered Gesture Acquisition and Recognition for a Multimodal Conversational System. *Proceedings of the 5<sup>th</sup> International Conference on Multimodal Interfaces (ICMI)*, pages 147–150, Vancouver, B. C., Canada, 5-7 November 2003.
- [37] AMI Project  
<http://www.amiproject.org>
- [38] AMI Corpus - The AMI Meeting Corpus  
<http://corpus.amiproject.org>
- [39] Chen, F., E. H. C. Choi and N. Wang. Exploiting Speech-Gesture Correlation in Multimodal Interaction. In *Proceedings of the 12<sup>th</sup> International Conference on Human-Computer Interaction*, LNCS4552, pages 23-30, Beijing, China, 22-27 July 2007.
- [40] Gundel, J. K., N. Hedberg and R. Zacharski. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, Volume 69, No. 2, pages 274–307, June 1993.
- [41] Kehler, A.. Cognition Status and Form of Reference in Multimodal Human-Computer Interaction. In *Proceedings of the 17<sup>th</sup> National Conference on Artificial Intelligence and 12<sup>th</sup> Conference on Innovative Applications of Artificial Intelligence (AAAI)*, pages 685–690, Austin, USA, 30 July-1 August 2000.
- [42] Oviatt, S., R. Coulston and R. Lunsford. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In *Proceedings of International Conference on Multimodal Interfaces (ICMI)*, pages 129–136, State College, PA, USA, 13-15 October 2004.
- [43] Coutaz, J., L. Nigay, D. Salber, A. Blandford, J. May and R. M. Young. Four Easy Pieces for Assessing the Usability of Multimodal Interaction:

- The CARE Properties. In *Proceedings of the IFIP TC13 5<sup>th</sup> International Conference on Human-Computer Interaction (INTERACT)*, pages 115–120, Lillehammer, Norway, 25-29 June 1995.
- [44] Corradini, A., M. Mehta, N. O. Bernsen, J. C. Martin and S. Abrilian. Multimodal Input Fusion in Human-Computer Interaction on the example of the on-going NICE project. In *Proceedings of the NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, 18-29 August 2003.
- [45] Nigay, L. and J. Coutaz. A Generic Platform for Addressing the Multimodal Challenge. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 98–105, Denver, USA, 7-11 May 1995.
- [46] Johnston, M., P. Cohen, D. McGee, S. Oviatt, J. Pittman and I. Smith. Unification-based Multimodal Integration. In *Proceedings of the 35<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 8<sup>th</sup> Conference of the European Chapter of the Association for Computational Linguistics (COLING-ACL)*, pages 281–288, Madrid, Spain, 1997.
- [47] Johnston, M.. Unification-based Multimodal Parsing. In *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 17<sup>th</sup> International Conference on Computational Linguistics (COLING-ACL)*, Volume 1, pages 624–630, Montreal, Canada, 1998.
- [48] Wu, L., S. Oviatt and P. Cohen. Multimodal Integration - A Statistical View. *IEEE Transactions on Multimedia*, Volume 1, Issue 4, pages 334–341, December 1999.
- [49] Wu, L., S. L. Oviatt and P. R. Cohen. From Members to Teams to Committee - A Robust Approach to Gestural and Multimodal Recognition.

- IEEE Transactions on Neural Network*, Volume 12, Issue 4, pages 972–982, 2002.
- [50] Johnston, M. and S. Bangalore. Finite-state Multimodal Parsing and Understanding. In *Proceedings of the 18<sup>th</sup> International Conference on Computational Linguistics (COLING)*, Volume 1, pages 369–375, Saarbrücken, Germany, 31 July-4 August 2000.
- [51] Chai, J. Y., P. Hong, M. X. Zhou and Z. Prasov. Optimization in Multimodal Interpretation. In *Proceedings of the 42<sup>nd</sup> Annual Meeting on Association for Computational Linguistics*, pages 1–8, Barcelona, Spain, 21–26 July 2004.
- [52] Qu, S. and J. Chai. Saliency Modeling based on Non-verbal Modalities for Spoken Language Understanding. In *Proceedings of the 8<sup>th</sup> International Conference on Multimodal Interfaces (ICMI)*, pages 193–200, Banff, Canada, 2-4 November 2006.
- [53] Rudnicky, A. and A. Hauptmann. Models for Evaluating Interaction Protocols in Speech Recognition. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems: Reaching through Technology*, pages 285–291, New Orleans, USA, 27 April-2 May 1991.
- [54] Hernández, Á., B. López, D. Díaz, R. Fernández, L. Hernández and J. Caminero. A “Person” in the Interface: Effects on User Perceptions of Multibiometrics. In *Proceedings of the Workshop on Embodied Language Processing*, pages 33–40, Prague, Czech Republic, 28 June 2007.
- [55] Sweeney, M., M. Maguire and B. Shackel. Evaluating User-Computer Interaction: A Framework. In *International Journal of Man-Machine Studies*, Volume 38, Issue 4, pages 689–711, April 1993.

- [56] Hui, P. Y. and H. Meng. Joint Interpretation of Input Speech and Pen Gestures for Multimodal Human-Computer Interaction. In *Proceedings of the 9<sup>th</sup> International Conference on Spoken Language Processing (Inter-speech 2006 - ICSLP)*, pages 1197-1200, Pittsburgh, USA, 17-21 September 2006.
- [57] Huang, X., A. Acero, C. Chelba, L. Deng, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Wang and Y. Wang. MiPad: A Next Generation PDA Prototype. In *the Proceedings of the 6<sup>th</sup> International Conference on Spoken Language Processing (ICSLP)*, Volume 3, pages 33-36, Beijing, China, 16-20 October 2000.
- [58] WordNet  
<http://wordnet.princeton.edu/>
- [59] Xiao, B., C. Girand and S. Oviatt. Multimodal Integration Patterns in Children. *Proceedings of the 7<sup>th</sup> International Conference on Spoken Language Processing (ICSLP)*, pages 629-632, Denver, Colorado, USA, 16-20 September 2002.
- [60] Xiao, B., R. Lunsford, R. Coulston, M. Wesson and S. Oviatt. Modeling Multimodal Integration Patterns and Performance in Seniors: Towards Adaptive Processing of Individual Differences. *Proceedings of the 5<sup>th</sup> International Conference on Multimodal Interfaces (ICMI)*, pages 265-272, Vancouver, B. C., Canada, 5-7 November 2003.
- [61] Bers, J., S. Miller and J. Makhoul. Designing Conversational Interfaces with Multimodal Interaction. In *Proceedings of DARPA Workshop on Broadcast News Understanding Systems*, pages 319-321, Lansdowne, VA, February 1998.

- [62] Brown, P. F., S. A. D. Pietra, V. J. D. Pietra and R. L. Mercer. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, Volume 19, pages 263–311, Issue 2, 1993.
- [63] Hui, P. Y., Z. Zhou and H. Meng. Complementarity and Redundancy in Multimodal User Inputs with Speech and Pen Gestures. In *Proceedings of Interspeech*, pages 2205–2208, Antwerp, Belgium, 27-31 August 2007.
- [64] Clarkson, P. and R. Rosenfeld. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of the 5<sup>th</sup> European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2707–2710, Rhodes, Greece, 22-25 September 1997.
- [65] Goodman, J. T.. A Bit of Progress in Language Modeling: Extended Version. *Technical Report MSR-TR-2001-72*, Microsoft Research, 2001.
- [66] Oviatt, S.. Multimodal System Processing in Mobile Environments. In *Proceedings of the 13<sup>th</sup> annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 21–30, San Diego, USA, 5-8 November 2000.
- [67] Chang, E., J. Zhou, S. Di, C. Huang and K. F. Lee. Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones. In *Proceedings of the 6<sup>th</sup> International Conference on Spoken Language Processing (ICSLP)*, pages 983–986, Beijing, China, 16-20 October, 2000.
- [68] HTK Speech Recognition Toolkit  
<http://htk.eng.cam.ac.uk/>
- [69] Reddy, G., P. Y. Hui and H. Meng. Pen Gesture Classification for Multimodal Human-Computer Interaction. *Internship Report MSCUJL-2007-03*, CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies (previously known as “Microsoft-CUHK Joint

- Laboratory for Human-centric Computing and Interface Technologies”), July 2007.
- [70] Oviatt, S.. Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 576–583, Pittsburgh, Pennsylvania, USA, 15-20 May 1999.
- [71] Meng, H., W. Lam and C. Wai. To Believe is to Understand. In *Proceedings of the 6<sup>th</sup> European Conference on Speech Communication and Technology (EUROSPEECH)*, Volume 5, pages 2015–2018, Budapest, Hungary, 5-9 September 1999.
- [72] Chan, S. F. and H. Meng. Interdependencies among Dialog Acts, Task Goals and Discourse Inheritance in Mixed-Initiative Dialog. In *Proceedings of the 2<sup>nd</sup> International Conference on Human Language Technology Research (HLT)*, pages 197–202, San Diego, USA, 24-27 March 2002.
- [73] Hui, P. Y. and H. Meng. Cross-modality Semantic Integration with Hypothesis Rescoring for Robust Interpretation of Multimodal User interactions. *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 17, Issue 3, pages 486–500, March 2009.
- [74] Naptali, W., M. Tsuchiya and S. Nakagawa. Word Co-occurrence Matrix and Context Dependent Class in LSA based Language Model for Speech Recognition. *International Journal of Computers*, Issue 1, Volume 3, pages 85–95, 2009.
- [75] Song, W. and S. C. Park. A Novel Document Clustering Model Based on Latent Semantic Analysis. In *Proceedings of the 3<sup>rd</sup> International Conference on Semantics, Knowledge and Grid (SKG)*, pages 539–542, Xian, China, 29-31 October 2007.

- [76] Chen, B.. Word Topic Models for Spoken Document Retrieval and Transcription. In *ACM Transactions on Asian Language Information Processing (TALIP)*, Volume 18, Issue. 1, No. 2, pages 2:1–2:27, 2009.
- [77] Lee, J. H., S. Park, C. M. Ahn and D. Kim. Automatic Generic Document Summarization based on Non-negative Matrix Factorization. *International Journal on Information Processing and Management*, Volume 45, Issue 1, pages 20–34, January 2009.
- [78] Bellegarda, J. R.. Latent Semantic Mapping: Principles and Applications. *Synthesis Lectures on Speech and Audio Processing*, Volume 3, No. 1, 2007.
- [79] Salton, G. and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *An International Journal on Information Processing and Management*, Volume 24, Issue 5, pages 513–523, 1998.
- [80] Hofmann, T.. Probabilistic Latent Semantic Analysis. In *Proceedings of the 15<sup>th</sup> Conference on Uncertainty in Artificial Intelligence (UAI)*, Stockholm, Sweden, 30 July-1 August 1999.
- [81] Blei, D. M., A. Y. Ng and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Volume 3, pages 993–1022, January 2003.
- [82] Salton, G. and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hall, New York, New Jersey, USA, 1983.
- [83] van Rijsbergen, C. J.. *Information Retrieval*. Butterworths, London, U.K., 1979.
- [84] Hui, P. Y., W. K. Lo and Helen Meng. *Usage Patterns and Latent Semantic Analyses for Task Goal Inference of Multimodal User Interactions*. In *Proceedings of International Conference on Intelligent User Interfaces (IUI)*, pages 129–139, Hong Kong, 7-10 February 2010.



- [85] Meng, H. and D. Li. Multilingual Spoken Dialog Systems. *Multilingual Speech Processing*, T. Schultz, K. Kirchhoff (Eds.), Academic Press, pages 399–443, 2006.
- [86] Wu, Z. Y., H. Meng, H. Ning and C. F. Tse. A Corpus-based Approach for Cooperative Response Generation in a Dialog System. In *Proceedings of the 5<sup>th</sup> International Symposium of Chinese Spoken Language Processing (ISCSLP)*, pages 614–626, Singapore, 13-16 December 2006.