CITYBROWSER II: A MULTIMODAL RESTAURANT GUIDE IN MANDARIN

Jingjing Liu, Yushi Xu, Stephanie Seneff, and Victor Zue

Spoken Language Systems Group, MIT Computer Science & Artificial Intelligence Laboratory

ABSTRACT

In this paper we present a conversational dialogue system, CityBrowser II, which allows users to inquire about information about restaurants in Mandarin. Developed in the Galaxy infrastructure with a common, language-independent semantic representation, CityBrowser integrates portability and scalability. By inheriting the infrastructure and main language understanding/generation components from its English predecessor, CityBrowser can easily be transformed to a Mandarin language environment. This paper describes our system implementation, focusing on the language-specific modifications to the original English system. We show that our language-independent yet scalable system infrastructure makes multilingualism a promising task.

Index Terms— Mandarin dialogue systems, language-independent infrastructure, portability

1. INTRODUCTION

For nearly two decades, our group has been conducting research leading to the development of multilingual spoken dialogue systems that combine multiple human language technologies (HLTs) to enable humans and machines to carry on a mixed-initiative conversation for interactive problem solving and information access [1, 2]. To ensure that these systems can easily be generalized to languages other than English, we have made two design choices. First, we assume that it is possible to extract a *common*, language-independent semantic representation from the languages of interest. Second, we require that each component in the system be as *language transparent* as possible to promote portability. As illustrated in Figure 1, the dialogue manager, discourse component, and the meaning representation are designed to be independent of the input or output language. Where language-dependent information is required, we have isolated it in the form of external models, tables, or rules.

In this paper, we describe our continuing efforts to explore the portability issue by developing generic HLT components and demonstrate their utility in a system framework. Specifically, we have derived a Mandarin dialogue system, CityBrower II, from its English predecessor, CityBrowser I [3, 4], our most recent multimodal spoken dialogue system that can provide information about restaurants, subway stations and major landmarks for ten major U.S. cities. An example scenario of CityBrowser II is shown in Table 1.

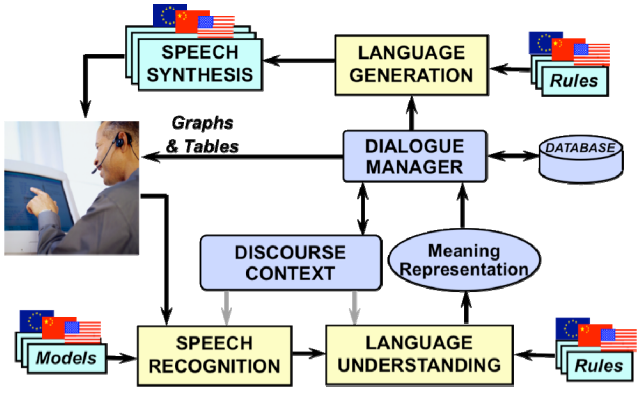

Figure 1. Architecture for multilingual spoken dialogue system.

Table 1. Example dialogue for CityBrowser II.

U1: 我想找一家日本料理餐馆
S1: 北京有 200 多家日本料理餐馆，它们大多集中在海淀区和朝阳区
U2: 海淀区知春路上有吗
S2: 知春路上有四家日本料理餐馆， 它们是…
U3: 哪一个比较便宜一点儿
S3: 一番日本料理比较便宜，它位于知春路 49 号地下一层
U4: 订餐电话是多少
S4: 电话号码是：88099658
U5: 清华附近呢
S5: 清华大学附近有 5 家比较便宜的日本料理餐馆，它们是...
U6: 有川菜吗
S6: 清华大学附近有 14 家川菜餐厅，它们大多数都比较便宜

The remainder of this paper is organized as follows: Section 2 describes the procedure of harvesting the database and scaling it to the target language environment; Section 3 describes the system architecture and gives a brief description of system functionality; Section 4 summarizes the system implementation and explains the language-relevant modifications to the original system; Section 5 provides an evaluation of the system and Section 6 summarizes our effort and points to future work.

2. HARVESTING THE DATABASE

Our first step in developing CityBrowser II is to replace the original English restaurant database with a Mandarin one that contains 6,533 restaurants in Beijing. The data are gleaned from crawling restaurant databases available on the web [1]. Restaurants are organized geographically around neighborhoods of 66 major universities.

The database is processed through a sequence of data clean-up steps: entity extraction, term tokenization, and proper noun expansion. We extract branch names for chain restaurants, and regularize district ("区", "县"), street, and street number from addresses. We also extend the ways cuisine and proper nouns can be expressed. Examples appear in Table 2.

Table 2. Examples of noun and proper noun generalization.

Cuisine		Neighborhood Name	
Original	**Alternatives**	**Original**	**Alternatives**
日本料理	日本菜, 日式	人大	人民大学
川菜	四川菜	清华大学	清华
广东菜	粤菜	北理工	北京理工大学

Chinese words are not separated by spaces, and thus require explicit tokenization. In spoken language people tend to refer to a named entity with a much shorter nickname, e.g., "全聚德" for "北京全聚德烤鸭店". To avoid the inflexibility of exact query matching in database look-up, we automatically generate a list of nicknames for each named entity. The generalization procedure works as follows: first, we optionally remove all aliases of the word "restaurant" from restaurant names, which include a set of more than 40 alternative forms, such as "酒楼" "酒店" "饭庄" "酒家". Then we remove locative nouns from the beginning of the named entity, such as "北京". Lastly, we remove the description of the cuisine, such as "烤鸭" and "火锅". The named entities can then be referred to by these derived variants.

We have also implemented a procedure to handle the restaurant hours, which requires more sophistication as there are various ways to express time intervals in Mandarin. We regularize different expressions into a uniform format and convert values in this format into Mandarin, such as from "10:30–1:40 6:00--0:00" to "上午十点半到下午一点四十, 晚上六点到凌晨十二点". An example of a cleaned-up restaurant entry is shown in Figure 2.

3. SYSTEM ARCHITECTURE

Like its predecessor, CityBrowser II is a web-based conversational system. A screenshot of the GUI appears in Figure 3, which consists of a dynamic web page centered around a Google map implemented with *Google Ditu* API[2], and a list of currently in-focus entities displayed to the right of the map. Audio communication is controlled via a Java applet embedded in the page which provides a push-to-talk button and endpointing. The GUI also supports zooming commands, such as "放大视图" and "在海淀区把地图缩小".

```
{q restaurant
    :rest_id 801
    :name "北京全聚德烤鸭店"
    :nickname "全聚德烤鸭店" "北京全聚德" "全聚德"
    :branch "北图分店"
    :address "海淀区白石桥路 37 号"
    :district "海淀"
    :street "白石桥路"
    :streetnum "37"
    :phone "68420527"
    :neighborhood("民族大学" "中央民族大学" "舞蹈学院"
"北京舞蹈学院")
    :cuisine "烤鸭"
    :hours "上午十点半到下午一点半, 下午四点半到晚上
九点半"
    :price_range "high" }
```

Figure 2. An example restaurant entry.

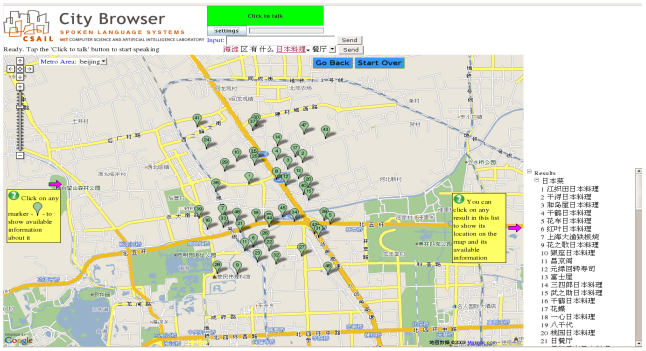

Figure 3. Screenshot of geographic GUI of CityBrowser II.

Both CityBrowser I and II utilize the open-source, Galaxy architecture [5] that provides a hub-based interaction among the HLT components, the database and the web-based GUI. The hub is responsible for frame-based message communication among the servers.

A typical dialogue turn works as follows: the input utterance from the user is sent from the GUI to the speech recognizer, which generates a list of *N*-best recognition hypotheses. The *N*-best hypotheses are parsed and rescored

[1] http://www.eatwell.com.cn/

[2] http://code.google.com/intl/zh-CN/apis/maps/. Since *Google Ditu* API presently only supports Geocoding at the district level and not the specific address, we have arbitrarily chosen to display each restaurant within the district by a randomized algorithm for the time being. This simulation will be omitted once Geocoding information becomes available for the restaurants.

by the TINA NLU component [6] and the best scoring hypothesis is transformed into a meaning representation, i.e., a linguistic frame [5]. The context resolution server [7] rewrites the linguistic frame with historical information and also stores it in history for the next turn. A list of key-value pairs is generated from the linguistic frame to represent the key concepts contained in the utterance. The dialogue manager consults the database based on these key-value pairs, retrieves matching entities, and composes a reply frame. A response utterance is finally generated from the reply frame and sent to the speech synthesizer, which generates audio and streams it to the GUI.

4. LANGUAGE-DEPENDENT COMPONENTS

In this section, we describe the system implementation, i.e., the procedure for transforming the language from English to Chinese, specifically in speech recognition and synthesis, as well as language understanding and response generation.

Speech Recognition & Synthesis:

The acoustic models for the speech recognition system, SUMMIT, are trained with a combination of two data corpora, "Yinhe" [8] and "MAT2000" [9], both of which contain Mandarin Chinese speech data from native speakers.

The class *n*-gram language model is trained [10, 11] by parsing a corpus using TINA [12]. Since we do not yet have a training corpus from users, we make use of the English corpus from CityBrowser I. We first create templates from translations into Mandarin of the English utterances. Then we automatically generate over 11,000 utterances in Mandarin from these translated templates. We train an individual recognizer for each category of domain-specific proper nouns (e.g., restaurant, neighborhood, street name). Then we replace these proper nouns in the training utterances with dynamic tags, such as "$dynrestaurant" for restaurant names, taking each category as a dynamic class. These dynamic-class recognizers are embedded in the utterance-level generic recognizer. Details of dynamic language modeling can be found in [11].

In addition, to scale to the specific restaurant domain, we extend a Mandarin vocabulary of the generic domain with 500 domain-specific entries, mostly Chinese restaurant names and street names. For the Mandarin synthesizer, we utilize a Mandarin text-to-speech system provided by the Industrial Technology Research Institute (ITRI).

Language Understanding:

To convert linguistic processing from English to Mandarin Chinese, we supply a generic-domain Mandarin grammar with a set of restaurant-domain class specifications. For example, we specify a Chinese cuisine category as an adjectival class covering 104 cuisines in Mandarin. Details of the Mandarin grammar and Chinese-specific context resolution approaches can be found in a companion paper [13]. As the semantic representation in Galaxy is language-independent, we can maintain the keys and generic values in English, while representing the values of database contents directly with Chinese characters, as shown in Table 3.

Table 3. Example of key-value list.

Input Utterance	<我想去一家在海淀区知春路上便宜一点儿的川菜餐馆>	
Key-Value List	**Unchanged**	**Translated**
	clause: request topic: restaurant price_range: cheap	cuisine: 川菜 district: 海淀 street: 知春路

Response Generation:

The dialogue manager remains unchanged from the English City Browser. It generates a response frame which needs to be converted into a well-formed Mandarin string, via GENESIS generation rules [14]. One key issue is word order differences between Mandarin Chinese and English. Other Chinese-specific challenges include word sense disambiguation. For example, a typical English query about the opening hours of a restaurant is: "Tell me the *hours* of the restaurant". In Chinese, however, simply translating "hours" to "小时" will bring up ambiguity. Our approach to disambiguation is to insert *flags* into templates.

The response generation rule for the template "do_hours" is as follows, where "!" represents a vocabulary lookup:

do_hours	:name :branch "的" ($:hours !hours)

This template sets up the $:hours flag, which is used by the vocabulary entry for "hours" to generate a context-dependent output string "营业时间" instead of its default string "小时".

5. EVALUATION

We performed a preliminary system evaluation by logging the interactions of 10 subjects with the system. All of the subjects are native speakers of Mandarin Chinese, 3 of whom are female speakers and 7 are male speakers. The subjects were recruited via email lists with an incentive of a $25 Amazon.com gift card.

We believe that Web-based multimodal systems should offer potential users a detailed context-dependent help mechanism to help them understand the scope of the domain. In the near future, we will incorporate a sophisticated "helper" capability in the Web page, which will provide users with utterance suggestions directly relevant to the current turn. Because this mechanism is not yet implemented, each subject went through a supervised practice session, where they were taught how to record and led through 15 single-turn interactions and three scenarios. Each subject was then given 10 scenario-based tasks with gradually increasing complexity. An example scenario is

shown in Table 4. A total of 836 recorded utterances led to recognition hypotheses from these subjects. Eight of the 10 subjects succeeded in all 10 tasks, and the other two subjects succeeded in 9 out of 10 tasks.

In addition to the interactions with the system, we also conducted a post-test survey on each subject. Each subject was given a paper-based survey to see if he/she finds the system easy to use, whether he/she thinks the restaurant-guide helpful, and whether he/she would like to recommend the system to their friends. Each question was evaluated on a Likert scale of 1 to 5 (1: very difficult, 5: very easy; 1: no recommendation, 5: highly recommend).

Figure 4 shows the annotations from each subject. Across all the subjects, the average perceived ease of use was 3.8; the average perceived helpfulness was 4.6; and the average recommendation was 4.0. Nine of the 10 subjects found the system very helpful, and 8 would recommend the system to their friends.

Table 4. An example scenario-based task.

You are a student at Tsinghua University. You have a friend visiting you, who likes Italian food very much. Your plan of the day is to first take him for lunch at a Japanese restaurant near school, then go window shopping at Wangfujing Street. After shopping in the afternoon, you will treat him at a fancy Italian restaurant nearby. You need to make reservations at both restaurants by phone, and you want to know exactly the locations of the restaurants.

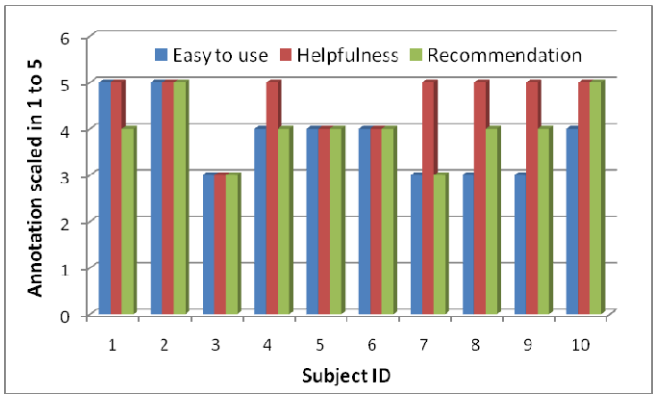

Figure 4. System evaluation by each subject.

Some subjects provided valuable feedback for system improvements, such as enabling look-up around landmarks like "Forbidden City" or "健翔桥", and providing restaurant rankings as well as comments from general users as additional information. We will integrate these functionalities in our future work.

6. SUMMARY

This paper describes our work in converting an English spoken dialogue system into a Mandarin one. By inheriting the portability and scalability of the Galaxy infrastructure and our language-independent framework, the system implementation requires only relatively minor changes to account for the differences between English and Mandarin Chinese. In the near future, we will continue to collect data from real users, and to expand the system to include more multimodal capabilities.

7. ACKNOWLEDGMENTS

This research is supported by Quanta Computers, Inc. through the T-Party project. We gratefully acknowledge ITRI for the use of their text-to-speech system. Thanks also goes to Alexander Gruenstein for providing tremendous help on many system and server issues.

8. REFERENCES

[1] C. Wang, S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi and V. Zue, "Muxing: A Telephone-Access Mandarin Conversational System," *Proc. ICSLP*, 2000.

[2] M. Nakano, Y. Minami, S. Seneff, T. J. Hazen, D. S. Cyphers, J. Glass, J. Polifroni, V. Zue, "Mokusei: A Telephone-based Japanese Conversational System in the Weather Domain," *Proc. EUROSPEECH*, 2001.

[3] A. Gruenstein and S. Seneff, "Releasing a Multimodal Dialogue System into the Wild: User Support Mechanisms," *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[4] A. Gruenstein, S. Seneff, and C. Wang, "Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Database," *Proc. Interspeech*, 2006.

[5] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," *Proc. ICSLP*, 1998.

[6] S. Seneff, "TINA: a natural language system for spoken language applications", *Computational Linguistics*, Vol. 18, No. 1, pp. 61-86, 1992.

[7] E. Filisko and S. Seneff, "A context resolution server for the Galaxy conversational systems," *Proc. EUROSPEECH*, 2003.

[8] C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff and V. Zue, "YINHE: A Mandarin Chinese Version of the Galaxy System," *Proc. Eurospeech 97*, *pp. 351-354*, 1997.

[9] H. C. Wang, F. Seide, C.Y. Tseng, L. S. Lee, "MAT2000-Design, collection, and validation of a Mandarin 2000-speaker telephone speech database," *Proc. ICSLP*, *pp. 460-463*, 2000.

[10] G. Chung, S. Seneff, C. Wang, and L. Hetherington, "A dynamic vocabulary spoken dialogue interface," *Proc. of INTERSPEECH*, 2004.

[11] A. Gruenstein and S. Seneff, "Context-Sensitive Language Modeling for Large Sets of Proper Nouns in Multimodal Dialogue Systems," *Proc. IEEE/ACL 2006 Workshop on Spoken Language Technology*, 2006.

[12] S. Seneff, C. Wang, and T. J. Hazen, "Automatic induction of n-gram language models from a natural language grammar," *Proc. EUROSPEECH*, 2003.

[13] Y. Xu, J. Liu, S. Seneff, "Mandarin Language Understanding in Dialogue Context," *ISCSLP*, 2008.

[14] L. Baptist and S. Seneff, "Genesis-II: A versatile system for language generation in conversational system applications," *Proc. ICSLP*, 2000.