YINHE: A MANDARIN CHINESE VERSION OF THE GALAXY SYSTEM[1]

Chao Wang, James Glass, Helen Meng, Joe Polifroni, Stephanie Seneff, and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{wangc, jrg, hmmeng, joe, seneff, zue}@sls.lcs.mit.edu

ABSTRACT

The GALAXY system is a human-computer conversational system providing a spoken language interface for accessing on-line information. It was initially implemented for English in travel-related domains, including air travel, local city navigation, and weather. We began an effort to develop multilingual systems within the framework of GALAXY several years ago. This paper describes our recent work on porting the system to Mandarin Chinese, including speech recognition, language understanding, and language generation components. Overall, the system produced reasonable responses nearly 70% of the time for spontaneous test data collected in a wizard environment.

1. INTRODUCTION

The GALAXY system is a client/server architecture for computer conversational systems [1]. In designing GALAXY, we drew heavily on experience gained in the development of GALAXY's predecessor, VOYAGER [2]. VOYAGER was not initially designed to easily support multiple languages, but through a trial-and-error process that involved several steps of redesign, we eventually developed a version of VOYAGER that could support three languages interchangeably – English, Italian, and Japanese [3]. The lessons learned from this exercise were carried over into the initial design of GALAXY, such that we believed it would be considerably more straightforward to port GALAXY to other languages besides English.

GALAXY is also a significantly more complex domain than was VOYAGER. It has *three* separate subdomains. One is a city guide similar to VOYAGER, except with a much larger set of known establishments available from an on-line Yellow Pages provided by NYNEX. In addition, GALAXY can also answer questions about (or make reservations for) flights worldwide from the American Airlines Sabre reservations system, and can give world-wide weather information obtained from the Web. The user can freely move from one domain to another in the course of a single conversation. Figure 1 shows the architectural plan of GALAXY. The user interface or *client* is quite lightweight and communicate with the user via speech and graphics, and with all of the various servers via the system's *hub*. The hub accesses the recognizer server, the NL server, and the domain servers to carry out its tasks. In general, there would be a synthesizer server as well, but we have not yet obtained one for Mandarin.

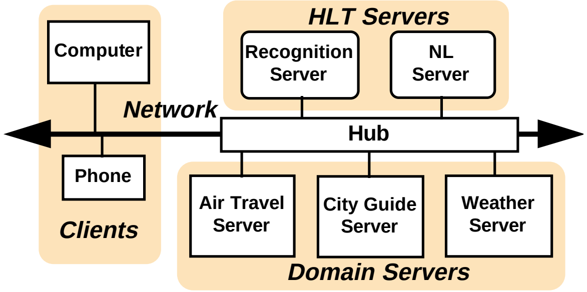


Figure 1: Block diagram of GALAXY architecture.

This paper concerns the development of YINHE[2], a Mandarin Chinese version of our GALAXY system. The user communicates with the system in Mandarin, and the system displays responses in Chinese ideography, along with maps, etc., as shown in Figure 2. We designed all parts of the system such that porting to a new language should involve only the replacement of external linguistic rules, acoustic models, language models, and vocabularies in the recognizer and NL servers. The system represents meaning in a hierarchical semantic frame format. The discourse component, as well as all of the domain servers, communicate with the GALAXY hub via these semantic frames, such that these components are all transparent to either the input or the output languages (which can differ). Data collection, however, remains a significant and time consuming language-specific effort, which is absolutely necessary for the development of a high-performance system.

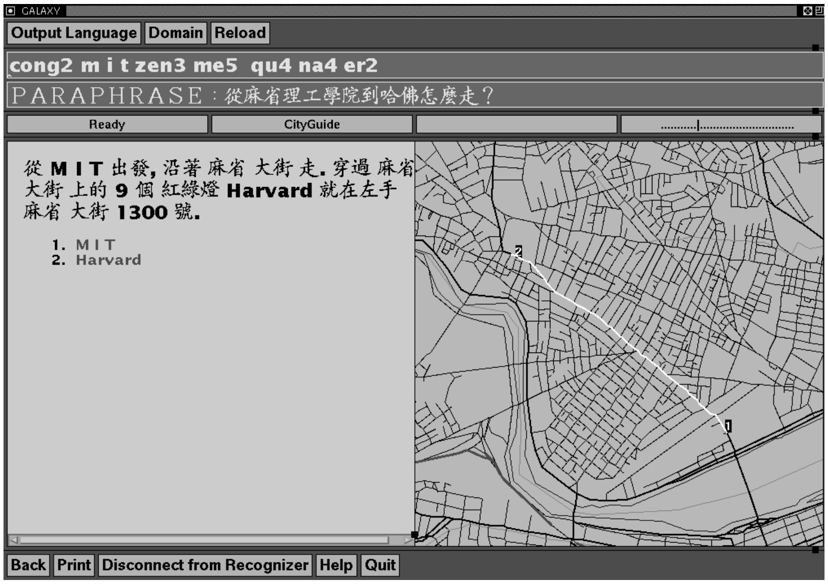
Overall, we consider the exercise of porting GALAXY to Mandarin to be a success. We now have a system which achieves comparable performance to its English counterpart. In the following sections, we will first describe our data collection effort in more detail. We will then discuss the particular issues that came up in porting the recognizer, the understanding component, and the generation component to Mandarin. Finally, we will give some evaluation results separately for the recognizer and the understanding components, and overall for the two components working in conjunction. We will conclude with a summary and a discussion of our future plans.

2. DATA COLLECTION

Both read and spontaneous speech have been collected from native speakers of Mandarin Chinese. Spontaneous

[1] This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center.

[2] Yinhe, or “silver river,” corresponds to “Galaxy” in Mandarin Chinese.



cong2 m i t zen3 me5 qu4 na4 er2

PARAPHRASE：從麻省理工學院到哈佛怎麼走？

從 MIT 出發, 沿著 麻省 大街 走. 穿過 麻省 大街 上的 9 個 紅綠燈 Harvard 就在左手 麻省 大街 1300 號.

1. MIT
2. Harvard

Figure 2: An example of a dialogue exchange between YINHE and a user. Subsequent to an inquiry about Harvard University, the user is asking for directions there from MIT. The system shows the path on the map and gives verbal directions, using discourse to infer Harvard as the referent for "there".

speech data were collected using a simulated environment based on the existing English GALAXY system. Each subject talked to the system in Mandarin Chinese. A bilingual wizard then translated each spoken utterance from Chinese into an equivalent English sentence and typed it into the system. The system then processed the query and displayed the response back to the subject in Chinese. In this way, the subject feels that he/she is interacting with a complete Chinese system, and the conversation is like that in the real application. The data were used for training both acoustic and language models for recognition, and deriving and training a grammar for language understanding.

In addition, a significant amount of read speech data was collected through our Web data collection facility [4]. For each subject, 50 sentences within the GALAXY domain were displayed in Chinese characters through the Web page. The subject was prompted over the telephone to read each utterance in turn. In this way, we could obtain good coverage of different handsets and lines, because the callers were randomly distributed. It is easier to collect read data in large amounts, and they are extremely valuable for acoustic training, due to the phone-line diversity.

We use pinyin, enhanced with tones, for Chinese representation in our transcription to simplify the input task. Homophones that are the same in both tones and base-syllables are indistinguishable in the pinyin representation. We determined however that this ambiguity could be resolved by the language understanding component and would not affect system performance. Chinese is distinct from English in that boundaries between words are not specified in the ideographic written form. We had the option of combining the syllables of pinyin into distinct word units. However, we decided not to tokenize the utterances into word sequences in the original transcription, because it is not always obvious even for native speakers what constitutes a word, and the selection of words would be likely to change during the development process. The sentences were later segmented into word sequences using a semi-automatic tokenization procedure based on a predefined vocabulary. Time-aligned phonetic transcriptions were not provided due to the tremendous effort required; instead, they were derived using a forced alignment procedure during the training process.

SET	TRAIN	DEV	TEST
No. of utts.	6,457	500	274
Type of utts.	Spon. and read		Spon.
No. of speakers	93		6
Words per utt.	8.3	8.5	8.0

Table 1: Summary of the corpus.

To date, we have collected about 3,100 spontaneous utterances from 64 speakers, and 4,200 read utterances from about 90 speakers, among whom 55 also participated in the spontaneous data collection. These speakers are from more than 15 provinces of China, which gives us a good coverage of various dialectal influences in the data. Speech data from 6 speakers were set aside to form a test set. The remaining utterances were divided randomly into a training set and a development set. A summary of the corpus is shown in Table 1.

3. SYSTEM DEVELOPMENT

3.1. Speech Recognition

The Chinese language is ideographic and tonal syllabic, in which each character represents a syllable with a particular tone, and one or more characters form a "word." Overall, there are about 416 base syllables and 5 lexical tones, including the "reduced" tone. In Chinese phonology, syllables are usually characterized by initials (syllable onset) and finals (the vowel nucleus and any final consonants). Syllable final consonants are restricted to /r/ and the nasals.

In YINHE, speech recognition is performed by the SUMMIT segment-based speech recognition system [5]. It would be advantageous to incorporate some kind of tone recognition into the framework. However, SUMMIT, as currently configured, does not have any capability for explicitly dealing with fundamental frequency, and it would also be difficult to incorporate scores provided at the syllable level. Thus, we have omitted tone recognition in the initial version of YINHE. We realize that this leads to a greater number of potential homophones, but most of these can be disambiguated at the parsing stage.

YINHE represents our first attempt at Mandarin speech recognition; thus there were no pre-existing local acoustic models or training corpora. Hence we depend critically on the domain-specific data we had collected. We also depended upon acoustic models that we had obtained for English phonemes, which were utilized as seed models for near-neighbor Mandarin models.

Vocabulary Specification

We went through several iterations to decide the actual vocabulary, mainly in making decisions about where to insert word boundaries in the syllable string. For example, the word "week" can be referred to in three different ways in Mandarin: "xing1 qi1," "li3 bai4," and "zhou1." The days of the week are formed by adding the numeric index of the day to this base word: "week 1," "week 2," etc. for "Monday," "Tuesday," etc. Thus an additional vocabulary of 21 words is needed to cover explicitly the days of the week. Of course, with the numeral attached, these words can be much better represented in a class bigram, which is a large benefit. At present, we are omitting the explicit knowledge of these words, recognizing that this is suboptimal in terms of the language model.

City names are prominent in GALAXY. Since they are world-wide, users are uncertain as to whether to refer to them in English or in Chinese. Hence we had to allow multiple entries for many of them, essentially an English and a Chinese equivalent. For example, "San Francisco" is referred to as both "san1 fan2 shi4" and "jiu4 jin1 shan1" (literally, Old Gold Mountain), so we have to maintain three distinct words for this one city. A similar problem exists for the place names in City Guide. We felt it would be difficult to cover all the odd pronunciations of restaurant names, etc. Therefore, we eliminated most of them from the vocabulary, thus encouraging the user to refer to them by index or by clicking.

The current vocabulary has about 1000 words, which is much smaller than that of the English system. About one quarter of the vocabulary are English, and each Chinese word has on average 2.1 characters.

Phonetic Models After some experiments with various sets of phonetic units, we finally settled on the simple choice of representing each syllable initial and final as an individual phonetic unit. We feel that our segment-based framework is particularly effective at capturing the dynamic nature of these multi-phoneme units, and the only problem was that we did not have very obvious English analogs for some of these units (such as "uan" and "iang") on which to seed. We were able to solve this problem by seeding any unusual finals on schwa, because of its inherent variability, along with an artificial reward during early iterations. This improves their chance of consuming the entire span of the syllable final during forced alignment, rather than giving part of it up to an undesirable insertion model or a neighboring syllable initial.

We also had some difficulty with the rich set of strong fricatives and affricates in Mandarin. Mandarin makes a distinction between /s/, /sh/, and a retroflexed /shr/. Similar distinctions are possible for the voiced and affricate counterparts. These phonemes are further complicated by the widespread regional differences among speakers. In Southern dialects, there is a tendency to lose the distinction of palatalization, so that instances of /s/ and /sh/ are nearly indistinguishable, Other dialects lose the distinction between /l/ and /n/ in syllable initial position. We initially tried handling these dialectal variations by phonological rules, but in the absence of hand-labeled data it became difficult to guarantee a correct realization in our training utterances. In the end we decided to let the models handle the variability through the Gaussian mixtures. We had an analogous problem with the Beijing dialect's tendency to retroflex the endings of certain syllables, which we treated in a similar way.

The English proper nouns are usually outside of the phonological and phonotactic structure of Mandarin. As a consequence, users often speak these names with a heavy accent, and it becomes problematic whether to build separate English phonetic models or to force these outliers into the nearest-neighbor Mandarin equivalent. For the most part we were able to share models, with the system being augmented with only a few phonemes particular to English, such as /v/ and /eh/. Thus, in some sense, we lexicalized the foreign accent for English, entering "New York" in the lexicon pronounced as "Niu Yok" and "South Boston" as "Saus Basteng."

3.2. Language Understanding

For language understanding we used the TINA system [6], which had been designed originally for English. Our approach to rule development was to determine the appropriate rules for each new Mandarin sentence by first parsing an English equivalent, and choosing, as much as possible, category names that paralleled the English equivalent. This minimized the effort involved in mapping the resulting parse tree to a semantic frame. While the temporal ordering of constituents is quite different for Mandarin than for English, the basic hierarchy of the phrase structure is usually very similar to that of English.

TINA has a trace mechanism to handle gaps that are prevalent for wh-questions in English (e.g., "[What street] is MIT on [trace]?"). In Mandarin, wh- words are not moved to the front of the sentence, and therefore these sentences are actually easier to accommodate than their English equivalents. Mandarin does however frequently utilize an analogous forward-movement strategy to topicalize certain constituents in a sentence. An example is given in Figure 3. Such sentences were well-matched to TINA's trace mechanism, which produces a desirable frame containing "in Boston" as a predicate modifying "museums," but paraphrasing properly, with "Boston" in the topicalized initial position, due to the trace marker.

We were a little uncertain about what to do with the tokenization problem – whether to include the partial tokenization that takes place at the time of recognition, or to disregard it and reparse without the commitments that the recognizer had made. We finally decided to discard the recognizer tokenization, and rely instead on the grammar rules of TINA to retokenize, with the belief that the final result would be more reliable. Since the grammar is heavily constrained by semantic categories throughout the parse tree, it is usually able to reconstruct the correct

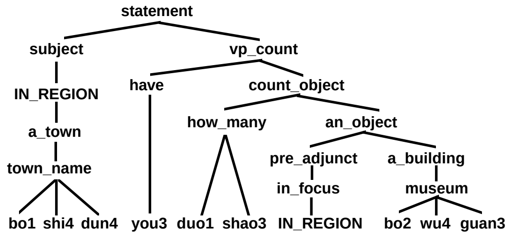


Figure 3: An example of long distance movement in Mandarin for topicalization, for the sentence, "Boston has how many museums?"

	No. of Utt.	WER	SER
Dev	500	9.1%	37.4%
Test	274	10.8%	39.1%

Table 2: Summary of the recognition performance.

tokenization of the sentence. We made a few exceptions to this rule, in cases where confusions with a common word could cause significant ambiguity. For instance, the first syllable of the word "jiu3 dian4," meaning literally "wine store," is a homophone for the word "jiu3" meaning "nine." Since numbers are prevalent in many different places in the grammar, we decided it was safer to commit to the whole word "wine store" up front, to expedite the parsing process. This effectively provides a one-syllable look-ahead to the parser.

3.3. Language Generation

We found that the process of generating correct paraphrases and responses in Mandarin Chinese was quite straightforward, and, for the most part, we were able to utilize our GENESIS framework without any changes [7]. One aspect of Mandarin that is quite different from English is the use of particles to accompany quantified nouns. These particles are analogous to "a *flock* of sheep" in English, except that they are far more pervasive in the language. Thus "four banks" becomes "four <particle> banks." Furthermore, the exact realization of the particle depends on the class of the noun, and there is a fairly large number of possibilities. For any language internal paraphrases (Mandarin => semantic frame => Mandarin) the particle can be parsed into the frame and reparaphrased intact. However for actual translation, the situation is problematic because complex context effects determine which particle to use under what circumstances. Similarly, Mandarin does not make obvious distinctions between singular and plural, which can be problematic when translating into English. Since YINHE is self-consistent with respect to language, these issues have been avoided, but we would like to be able to produce trans-lingual paraphrases that are also well-formed.

4. EVALUATION

Table 2 shows the speech recognition performance in terms of word error rate and sentence error rate on the development and test data. Table 3 shows the speech understanding performance on the test data. The 10-best entry gives the results obtained based on the parse selected automatically from a 10-best list. There are 20% of the queries which, if recognized perfectly, would still not be understood correctly. Thus the gap between speech input and text input is only 10 percentage points. Most of the sentences that fail to parse are outside of the domain of GALAXY or suffer from disfluencies which are beyond the limited robust parsing capabilities of YINHE.

	Parsed			Failed
	Perfect	**Acceptable**	**Wrong**	
1-best	62.41	6.93	2.56	28.10
10-best	69.71	8.39	5.48	16.42
Ortho.	80.29	2.92	0.73	16.06

Table 3: Speech understanding performance in percentages on the 274 spontaneous utterances in our test set.

5. SUMMARY AND FUTURE WORK

This paper has described our implementation of a Mandarin version of our GALAXY system. We feel that the success of this effort demonstrates the feasibility of our design aimed at accommodating multiple languages in a common framework.

While our system converses with the user completely in Mandarin, it often displays information it has obtained, for example from the Web, in English. Thus, if the user asks for the weather, it says, in Mandarin, "Here is the weather for Beijing," and then shows the weather report in English. We have actually begun the process of translating on-line weather reports from English to Mandarin, so that the information itself, and not just the remark about the information, will be provided to the user in their preferred language.

We plan to add a tone-recognition capability to our recognizer. This may require us to restructure the framework to accommodate explicit knowledge of syllable boundaries.

6. REFERENCES

[1] D. Goddeau, E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. Zue, "GALAXY: A Human-language Interface to On-line Travel Information," *ICSLP '94*, Yokohama, Japan, pp. 707–710.

[2] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "The VOYAGER Speech Understanding System: A Progress Report," *Proc. Second DARPA Speech and Natural Language Workshop*, Harwichport, MA, Oct. 1989, pp. 51–59.

[3] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual Spoken-language Understanding in the MIT VOYAGER System," *Speech Communications*, 17(1-2), 1995, pp. 1–19.

[4] E. Hurley, J. Polifroni, and J. Glass, "Telephone Data Collection Using the World Wide Web", *ICSLP '96*, Philadelphia, PA, pp. 1898–1901.

[5] J. Glass, J. Chang, and M. McCandless, "A Probabilistic Framework for Feature-Based Speech Recognition," *ICSLP '96*, Philadelphia, PA, pp. 2277–2280.

[6] S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, 18(1), 1992, pp. 61–86.

[7] J. Glass, J. Polifroni, and S. Seneff, "Multilingual Language Generation Across Multiple Domains," *ICSLP '94*, Yokohama, Japan, pp. 983–986.