Porting the Galaxy System to Mandarin Chinese

by

Chao Wang

B.S., Tsinghua University, Beijing, China, 1994

Submitted to the Department of Electrical Engineering
and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1997

Author .....
Department of Electrical Engineering
and Computer Science
June 11, 1997

Certified by .....
Stephanie Seneff
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Porting the Galaxy System to Mandarin Chinese

by

Chao Wang

Submitted to the Department of Electrical Engineering
and Computer Science
on June 11, 1997 in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

GALAXY is a human-computer conversational system that provides 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. Efforts were started to develop multilingual systems within the framework of GALAXY several years ago. This thesis focuses on developing the Mandarin Chinese version of the GALAXY system, including *speech recognition*, *language understanding* and *language generation* components. Large amounts of Mandarin speech data have been collected from native speakers to derive linguistic rules, acoustic models, language models and vocabularies for Chinese. Comparisons between the Chinese and English languages have been made in the context of system implementation. Some issues that are specific for Chinese have been addressed, to make the system core more language independent.

Overall, the system produced reasonable responses nearly 70% of the time for spontaneous Mandarin test data collected in a "wizard" mode, a performance that is comparable to that of its English counterpart. This demonstrates the feasibility of the design of GALAXY aimed at accommodating multiple languages in a common framework.

Thesis Supervisor: Stephanie Seneff

Title: Principal Research Scientist

Acknowledgments

I wish to express my deepest gratitude to Stephanie Seneff, my thesis advisor, for the advice, support, encouragement she has extended to me over the past 15 months; it has been a great pleasure to learn from her and work with her closely. In addition, I wish to thank her for reading and editing this thesis. Without her, this thesis would not have been possible.

I thank everyone in the Spoken Language System group: Victor Zue, for giving me the opportunity to become a member of the SLS group, and stimulating an excellent research environment that is productive and fun; Jim Glass, for his help in setting up the SUMMIT recognizer for Mandarin Chinese; Helen Meng, for her many helpful suggestions and discussions; Joe Polifroni, Christine Pao, and Ed Hurley, for their assistance in data collection and maintaining the system; Vicky Palay and Sally Lee, for keeping things in order; Jane, for being such a wonderful officemate and helping me in many ways; TJ, Grace, Jim, Ray, Kenney, Drew, Sri, Giovanni, Ben, Mike, Michelle, Aarati, Jon and everyone else, for the answers and suggestions they provided and the comfortable and friendly environment they created.

I would also like to thank my Chinese friends and fellows deeply for contributing more than 7,000 Mandarin utterances to support this work. Without these data, this work would not have been possible.

Finally, I wish to thank my family, Jinlei, Mom, Dad and my brother, for their enduring love and support.

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

Contents

1 Introduction **10**

1.1 Previous Research 11

1.1.1 Conversational Systems 11

1.1.2 Multilingual Translation Projects 11

1.1.3 Multilingual Conversational Systems 12

1.2 Background 13

1.2.1 General Description of Galaxy 13

1.2.2 Multilingual Approach [38] 14

1.2.3 Development Procedure 16

1.3 Objectives 18

1.4 Thesis Outline 18

2 Mandarin Speech Recognition **20**

2.1 Corpus 21

2.1.1 Data Collection 21

2.1.2 Data Analysis 23

2.2 The Summit Speech Recognition System 23

2.2.1 Signal Processing 24

2.2.2 Segmentation 24

2.2.3 Feature Extraction 24

2.2.4 Acoustic Modeling 26

2.2.5 Search 26

2.2.6 Iterative Training Process 26

2.3 Issues in Mandarin Speech Recognition 27
2.3.1 Vocabulary 27
2.3.2 Acoustic Modeling 28
2.3.3 Dialectal Variations 33
2.3.4 Tokenization 35
2.3.5 Homophones 37
2.4 Summary 39

3 Natural Language Understanding 40
3.1 Grammar Rules 41
3.1.1 Writing Grammar Rules 41
3.1.2 Special Issues in Parsing Chinese 44
3.2 Special Processing 47
3.2.1 Long-distance Movement 48
3.2.2 Keyword Mapping 50
3.2.3 City-state Decoding Algorithm 52
3.3 Performance Analysis 54
3.4 Interface between Recognizer and Parser 56
3.5 Summary 57

4 Natural Language Generation 59
4.1 Linguistic Rules for Generation 60
4.1.1 Message Templates 60
4.1.2 Lexicon 63
4.1.3 Rewrite Rules 64
4.2 Generation Processes 64
4.2.1 Response Generation 64
4.2.2 Paraphrase Generation 65
4.3 Issues in Chinese Generation 67
4.3.1 Topic Movement 67
4.3.2 Particles 68

4.3.3 Difficulties for Trans-lingual Paraphrasing 70
4.4 Summary 70

5 Evaluation **72**
5.1 Evaluation Methodology 72
5.1.1 Data 72
5.1.2 Methodology 73
5.2 Results and Analysis 75
5.2.1 Speech recognition 75
5.2.2 Language Understanding 75
5.2.3 Speech Understanding 76
5.2.4 Language Generation 77

6 Summary and Future Work **78**
6.1 Summary 78
6.2 Future Work 81

List of Figures

1-1 Example queries for each subdomain of GALAXY 13

1-2 User interface of the GALAXY system (Mandarin Chinese version) . . 14

1-3 Architecture of MIT's multilingual conversational system 15

1-4 Development procedure for porting to a new language 16

2-1 Block diagram of the SUMMIT speech recognizer 24

2-2 The recognition of an example Mandarin utterance. Shown from top to bottom are the waveform, spectrogram, segment hypothesis network (associated with scores), best-paths (both phonetic and orthographic), and forced-paths (both phonetic and orthographic). The small inset window shows the scores of different phone hypotheses for the segment between the two vertical time markers. 25

2-3 Decomposition of syllable structure in Mandarin Chinese 29

2-4 Recognition performance with varying maximum number of mixtures 32

2-5 Summary and examples of tokenization errors 36

2-6 Example of different ways of tokenization 36

2-7 Example of type 1 homophones 38

2-8 Example of type 2 homophones 38

2-9 Algorithm to correct type 2 homophones 39

3-1 Parse tree and semantic frame for the Chinese sentence "Show from New York to Beijing +s flight." 42

3-2 Parse tree and semantic frame for the corresponding English sentence "Show me flights from New York to Beijing." 43

3-3 Example parse tree illustrating tokenization and homophone disambiguation 45
3-4 A left-recursive grammar and the non-recursive implementation . . . 47
3-5 Examples of Chinese sentences without long-distance movement . . . 48
3-6 Examples of Chinese sentences with long-distance movement 49
3-7 Parse tree and semantic frame for an example Chinese sentence with long-distance movement 49
3-8 An example semantic frame without keyword mapping 51
3-9 Example sentences to illustrate context-dependent keyword translation 51
3-10 Examples of semantic frames without keyword translation 52
3-11 Examples of semantic frames with keyword translation 52
3-12 Examples of out-of-domain sentences 54
3-13 Examples of disfluent sentences 54
3-14 Parsing coverage on recognizer N-best hypotheses with varying N . . 56

4-1 An example of response generation for directions in the City Guide domain, which produces the Chinese equivalent of "Starting from MIT, follow the traffic on Massachusetts Avenue. Continue through 9 lights on Massachusetts Avenue. Harvard will be at 1300 Massachusetts Avenue on your left." 65
4-2 Semantic frame for the Chinese sentence "Show from New York fly to Beijing +s flight." 65
4-3 Semantic frame for an example Chinese sentence with long-distance movement 67

5-1 System architecture model for evaluation 73
5-2 Examples of test set sentences that failed to parse 76

6-1 Parse tree illustrating the syllabic feature of Chinese 81

List of Tables

2-1 Summary of the corpus 23

2-2 Chinese vocabulary for weekday names 27

2-3 Acoustic-phonetic symbols and example occurrences 30

2-4 Examples of dialectal variations 34

3-1 Summary of parsing coverage on the orthographies of the spontaneous training utterances 55

3-2 Quality analysis of parsed recognizer 10-best outputs on development set data 57

4-1 Example message templates for the city guide domain 62

4-2 Example message templates for the air travel domain 62

4-3 Example lexical entries for the city guide domain 63

4-4 Example lexical entries for the air travel domain 63

4-5 Message templates, lexical entries, and rewrite rules used in a long-distance movement example for language generation 69

5-1 Summary of recognition performance on test set data 75

5-2 Summary of language understanding performance on test set data . . 76

5-3 Summary of speech understanding performance on test set data . . . 76

Chapter 1

Introduction

As computers and information are becoming a more and more integrated part of our lives, the research involved in improving the human computer/information interaction interface is becoming increasingly prominent. A *speech* interface, in the user's *own language*, is highly desirable because of its naturalness, flexibility and efficiency [32].

GALAXY [17] is a human-computer conversational system developed in the SLS group at MIT that demonstrates a spoken language interface for accessing on-line information and services. The system tries to understand a user's speech queries, find out the answers, and deliver them back to the user in natural language as well as other formats. GALAXY was initially implemented for English in travel related domains, including air travel, local city navigation and weather information. Efforts have been made to develop multilingual conversational systems within the common framework of GALAXY [38]. This thesis focuses on developing the Mandarin Chinese version of the system, including speech recognition, language understanding, and language generation components. Large amounts of Mandarin speech data have been collected from native speakers to derive linguistic rules, acoustic models, language models and vocabularies for Chinese. Comparisons between the Chinese and English languages have been made in the context of system implementation. Issues that are specific for Chinese have been addressed, to make the system core more language independent.

1.1 Previous Research

With the advances in the automatic speech recognition area, many research efforts have been directed towards the development of spoken language systems for various applications since the late 1980's. In this section, we introduce some different types of spoken language systems that are related to this thesis.

1.1.1 Conversational Systems

Conversational systems are typically applied in information retrieval and interactive transactions, in which the user interacts with the computer in spoken language, with both of them being an active participant in the conversation. Several language-based technologies must be integrated to accomplish this task. On the input side, *speech recognition* must be combined with *natural language processing* in order for the computer to derive an understanding of the spoken input; on the output side, *language generation* capability must be developed in order for the system to communicate with the user also in natural language, verbally using a synthesizer. *Discourse management* also needs to be incorporated into the system to make the conversation natural and smooth.

Many human-computer conversational systems have appeared in the last few years, in which a user can typically carry on a spoken dialogue with a computer, within a narrow domain of expertise [3, 4, 8, 9, 14, 24, 26, 27, 28, 33]. For example, the WAXHOLM system [4] gives information on boat traffic in the Stockholm archipelago. There are other systems for accessing time schedules for trains [8, 9, 26]; the ATIS (Air Travel Information System) domain [27, 33]; urban navigation [3, 35]; automobile classifieds [24]; etc.

1.1.2 Multilingual Translation Projects

The goal for multilingual translation systems is to enable humans to communicate with one another in their native tongues. To achieve this goal, *speech recognition*, *machine translation*, and *speech synthesis* technologies must be combined.

There are several ongoing multilingual speech translation projects, usually under international collaborations among many research groups [2, 5, 22, 25, 31]. These systems typically involve human-human interactive tasks such as negotiation and scheduling. For example, the VERBMOBIL project [5] aims at developing a portable simultaneous speech-to-speech translation system for face-to-face negotiation dialogs, which currently supports Japanese-to-English and German-to-English translation; the JANUS system [22] developed at CMU facilitates translating spoken utterances from English or German to one of German, English or Japanese; the SLT (Spoken Language Translation) system [31] developed at ETRI translates Korean to English or Japanese for a travel planning task; etc.

1.1.3 Multilingual Conversational Systems

The GALAXY system is different from these systems in that it provides multilingual human-computer interfaces such that the information stored in the database can be accessed and received in multiple spoken languages. We believe that such systems are very useful since information is fast becoming globally accessible. We also suspect that this type of multilingual system may be easier to develop than the speech translation systems, because the system only needs to deal with the diversity of the conversation from the human side, and the topic of the conversation is usually focused in the application domain.

We have not found other similar systems reported in the literature so far, aside from GALAXY's predecessor VOYAGER [34, 35]. VOYAGER only has a city guide domain which is much simpler than the current city guide subdomain of GALAXY, and it was not initially designed to easily support multiple languages. Through a trial-and-error process that involved several steps of redesign, it eventually could support three language interchangeably – English, Japanese and Italian [10, 14, 15]. The lessons learned from this exercise were carried over into the initial design of GALAXY, such that it would be considerably more straightforward to port GALAXY to other languages besides English. A more detailed background about the GALAXY system is given in the following section.

1.2 Background

1.2.1 General Description of Galaxy

GALAXY provides a speech interface for accessing on-line information and services. It has *three* subdomains in its knowledge base: *city guide*, *air travel* and *weather*. The city guide domain knows about a large set of establishments in the greater Boston area, available from an on-line Yellow Page provided by NYNEX; the air travel domain can answer questions about (or make reservations for) domestic and international flights from the American Airlines Sabre reservations system; the weather domain can provide world-wide weather information derived from the Web. The user can freely move from one domain to another in the course of a single conversation. Figure 1-1 shows some example queries for each of the subdomains.

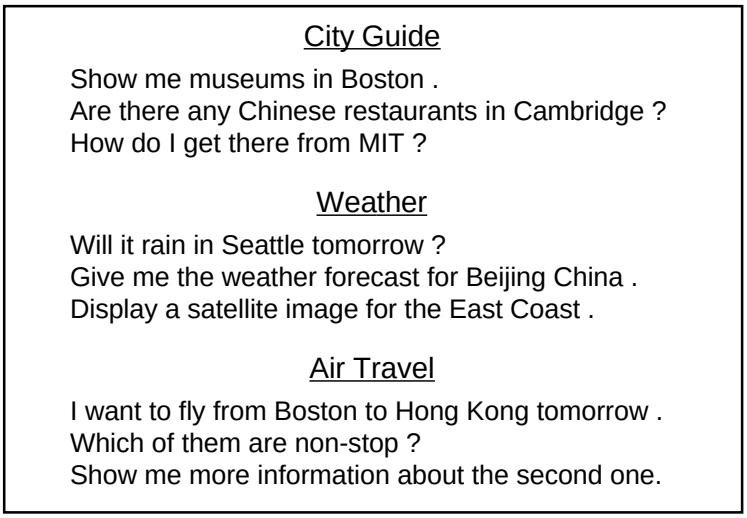


Figure 1-1: Example queries for each subdomain of GALAXY

GALAXY interfaces with the user through a *client* window and a *telephone*. Figure 1-2 shows the Mandarin Chinese version of the system as seen by the user. The pinyin sequence on the top of the window is the recognition output; the line below it shows a paraphrase (with discourse information incorporated) of the input utterance in Chinese ideography. In this example, the user is asking for directions *there* from

MIT after an inquiry about Harvard University. The system displays the path on the map and gives verbal directions in Chinese ideography, using discourse to infer Harvard as the referent for "there". Usually the system can also play back the verbal response through a synthesizer; however, we have not yet obtained a synthesizer for Mandarin Chinese.

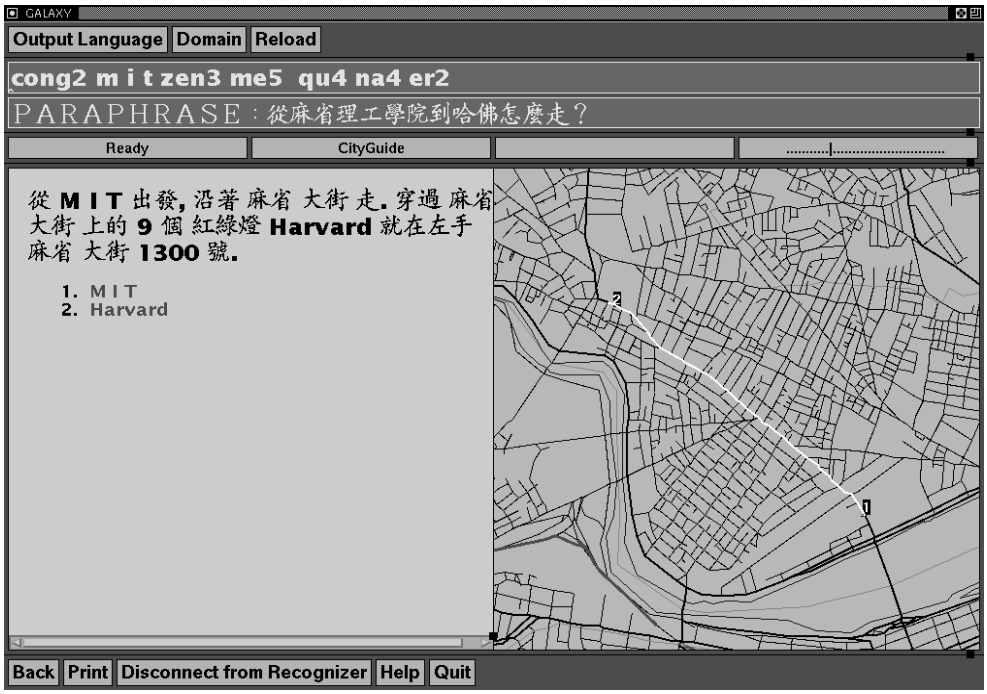


Figure 1-2: User interface of the GALAXY system (Mandarin Chinese version)

1.2.2 Multilingual Approach [38]

Figure 1-3 shows the architecture used by the GALAXY system, emphasizing its multilingual nature. Speech recognition is performed by the SUMMIT [13] segment-based, probabilistic, continuous speech recognition system. The *N*-best sentence hypotheses from SUMMIT are parsed by the TINA [29] natural language processing system, producing a meaning representation of the sentence called a *semantic frame*. The discourse component incorporates proper discourse information into the semantic frame to form a complete meaning representation. The semantic frame is then used by the

system manager to get appropriate information from the database and display the results back as tables and graphics. In practice, this involves consulting several domain servers; each specializing in a particular subdomain (eg., city guide). Natural language responses are also generated from the semantic frame by the GENESIS [16] language generation system. The verbal responses are played back to the user through an audio channel when a synthesizer for that language is available.

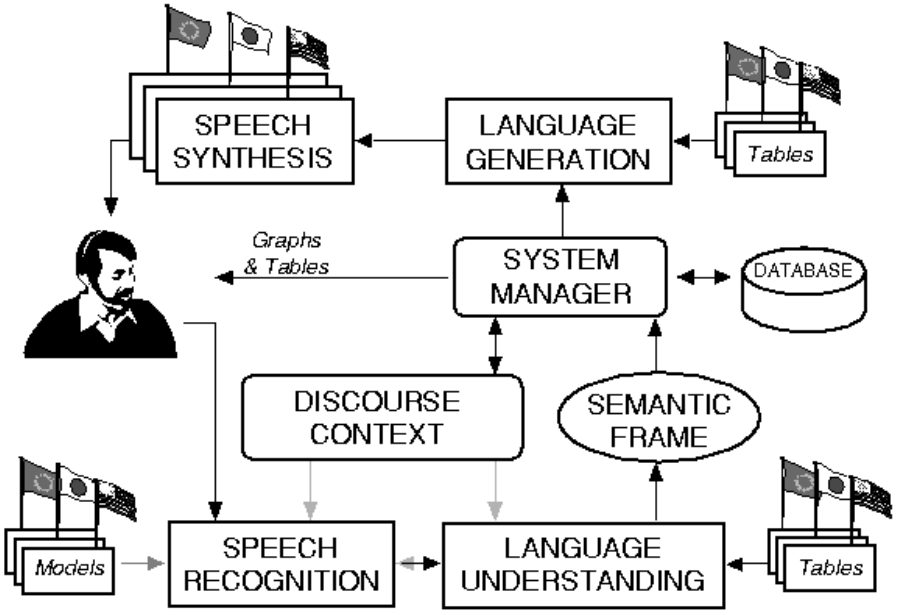


Figure 1-3: Architecture of MIT's multilingual conversational system

Our approach to developing multilingual conversational systems is based on the assumption that it is possible to extract a common, language-independent semantic representation for all the different languages, similar to the *interlingua* approach to machine translation [21]. We suspect that this assumption is highly probable to be valid in restricted domains, since the input queries will be goal-oriented and more focused. In addition, the semantic frame may not need to capture all the details in the input natural sentence. It suffices if the computer can use the semantic frame to get the correct information the user is seeking.

In order for the system to have multilingual capability, each component in the

system is designed to be as language transparent as possible. As shown in Figure 1-3, the semantic frame, system manager, discourse component, and database are all structured to be independent of the input or output language. Where language-dependent information is required, it is isolated in the form of external models, tables, or rules for the speech recognition, language understanding, and language generation components. So the task of porting to a new language should involve only adapting existing tables, models and rules without modification of the individual components. And the architecture of each component will slowly be generalized to be more language independent when new languages are incorporated. Details about the GALAXY system and each component can be found in [13, 16, 17, 29, 38].

1.2.3 Development Procedure

The procedure to port the GALAXY system to Mandarin Chinese is carried out according to the steps shown in Figure 1-4.

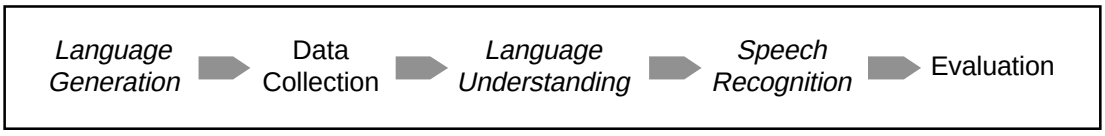


Figure 1-4: Development procedure for porting to a new language

Natural Language Generation

The first step is to generate the appropriate responses in Chinese from the semantic frames, which are derived from a set of English training sentences. This is done by providing a set of message templates, a lexicon, and a set of rewrite rules for Chinese for the GENESIS system. The purpose of developing language generation capability first is to obtain a simulated Chinese system environment for *wizard mode* data collection, and we are able to carry out this step first because no Chinese training data are needed at this stage.

Data Collection

Second, we construct a simulated Chinese system using the existing English system and the developed Chinese response generation ability, and collect sentences from native Chinese speakers who are brought in to interact with the system. The subject speaks to the system in Mandarin Chinese. I then translate the spoken utterance from Chinese into an equivalent English sentence and type it into the system. The system then processes the query and displays 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 can then be used for training the acoustic and language models for recognition, and deriving and training the grammar for language understanding.

Natural Language Understanding

Third, the collected sentences are used as training data to develop a grammar for Chinese and train the probabilities. Initially, each sentence is translated by hand into a list of rules invoked to parse it. After the grammar has built up a substantial knowledge of the language, many new sentences can be parsed automatically, or with minimal intervention to add new rules. The capabilities of the natural language component TINA are extended to handle the problems that arise from the special nature of the Chinese language.

Speech Recognition

Fourth, lexical items (with associated pronunciations), acoustic models, and language models for Mandarin Chinese are derived from the training sentences. The SUMMIT recognition engine can then use these parameters to bring up speech recognition ability for Mandarin Chinese.

Evaluation

Finally, the performance of the system must be evaluated using previously unseen data. And the capabilities of the system will be improved and refined as more training data are acquired.

1.3 Objectives

This thesis concerns porting the GALAXY conversational system to Mandarin Chinese. The primary objective of this thesis is to demonstrate the GALAXY system's feasibility as a multilingual architecture for human-computer interaction, especially as applied to a language that is substantially different from English. In particular, the thesis achieves three goals:

- Extend the GALAXY system to Mandarin Chinese, so that we can demonstrate a Mandarin speech interface for accessing on-line information.
- Compare the similarities and differences between the Chinese and English languages in the context of implementation of the GALAXY system, and generalize the architecture of the speech recognition, language understanding and language generation component to accommodate the special nature of the Chinese language.
- Study the issues in multilingual system development.

1.4 Thesis Outline

The remainder of the thesis contains five chapters. Chapter 2 covers the development of the Mandarin speech recognition system. The SUMMIT recognizer is configured properly for this purpose. Since lexical items, acoustic models and language models are derived from training data, the data collection effort is first described and an analysis of the data performed. After that, the basics about the SUMMIT system

are briefly introduced. Some issues that came up in porting the SUMMIT system to Mandarin Chinese are then described in detail and solutions proposed.

The natural language processing for Chinese is discussed in Chapter 3. An efficient way to port the grammar from English to Chinese is introduced. Some problems arising from the Chinese grammar are described, and solutions based on the existing system are proposed. Various special processing mechanisms in TINA are applied to process Chinese directly or with minor modifications. The interface between the recognizer and the parser is described, and some experimental results on development data are analyzed.

The language generation for Chinese is introduced in Chapter 4. The basic principles of the GENESIS system are introduced. A set of message templates, a lexicon, and a set of rewrite rules are developed for Chinese, and the processes for system response generation and paraphrasing input sentences are introduced. Some of the special language phenomena in Chinese are described and solutions proposed. Some difficulties for trans-lingual paraphrasing (translation) are pointed out.

Formal evaluations on unseen test data are presented in Chapter 5. Since the system manager, discourse component, and database are beyond the scope of this thesis, the evaluation methodology is designed to exclude their influences. The recognizer, the language understanding component, and the language generation component are evaluated separately; the overall evaluation results for the recognition and understanding components working in conjunction are also provided.

Chapter 6 summarizes the thesis and discusses future work.

Chapter 2

Mandarin Speech Recognition

The Chinese language is ideographic and tonal syllabic, in which each character is formed by strokes and pronounced as a mono-syllable with a particular tone. A Chinese character is usually a morpheme, and one or multiple characters constitute a word. Since a sentence is expressed as a sequence of characters without landmarks such as spaces between words, the interpretation of words could be ambiguous in many cases. Overall, there are about 416 base syllables and 5 lexical tones, including the "reduced" tone. Not all the tone-syllable combinations are legitimate, and the entire language consists of only about 1,345 unique tone-syllable pairs. Considering that there are about 6,000 commonly used characters in the language, it is very often the case that distinct Chinese characters may correspond to the same tonal syllable, forming homophones. It has previously been found that the vocal tract configurations are somewhat independent of tone production, which is primarily observed in the F_0 contour. Therefore, previous approaches generally process tone and base syllable recognition separately or totally ignore tone recognition (which leads to an even higher chance of homophones). Disambiguation of homophones is hoped to be solved through language modeling [11, 19, 23, 30].

For conversion from Mandarin speech to words, we use the SUMMIT probabilistic, feature-based speech recognition system [13], configured appropriately. The lexicon is constructed at the word level, and the acoustic features are modeled at the sub-syllabic level to reduce the number of models needed. A class bigram is used for

language modeling. The tone recognition is ignored, and we propose a method to use the language understanding system TINA to resolve the homophones. This chapter begins with a description of the corpus used to develop our system. Next, the basics about the SUMMIT system are introduced briefly. After that, we discuss some of the problems we have solved for Mandarin speech recognition in more detail. Formal evaluation on unseen test data will be given in Chapter 5.

2.1 Corpus

In order to obtain good recognition results within the application domain, it is necessary to acquire speech data specific to the domain to train acoustic, lexical and language models. Domain-specific data are also needed to derive a grammar and train the probabilities of the grammar for the language understanding system. In this section, we describe our data collection effort, and provide an analysis of the data.

2.1.1 Data Collection

Both read and spontaneous speech have been collected from native speakers of Mandarin Chinese. The spontaneous data are highly desirable because they better match the situation in real application and are suitable for all aspects of system training, but it is very time consuming to collect them in large quantity. On the other hand, read speech data can be collected in a more automated manner, and they are very useful for robust acoustic training.

Spontaneous Speech Data Collection

Spontaneous speech data were collected using a simulated environment based on the existing English GALAXY system. The subject talked to the system in Mandarin Chinese. I then served as a bilingual typist to translate the spoken utterances from Chinese into an equivalent English sentence and type 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 will be much like that in the real application. The data were used for training both the acoustic and language models for recognition, and deriving and training the grammar for language understanding.

Read Speech Data Collection

Read speech data are collected through our Web data collection facility [20]. For each visitor, 50 Chinese sentences typical of the application domains are displayed in ideography through the Web page, and the subject will be called at the phone number he or she has specified, and prompted to read each utterance in turn. We can get a good coverage of different telephone handsets and lines because the callers are random. Since the occurrences of read utterances are artificial, they are not suitable for training the probabilities for the language model and the grammar. But it is easier to collect them in large amounts, and they are very valuable for robust acoustic training because of the diversity in telephone handsets and lines, as well as the opportunity to guarantee coverage of rare words in the lexicon.

Transcription

The data were transcribed manually to form a corpus. We use tonal pinyin 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 will show that this ambiguity could be resolved by the language understanding component and would not affect system performance. We did not tokenize the utterances into word sequences in the transcription, because the definition of words is not obvious and might change during the development process. The sentences were later segmented into word sequences using a semi-automatic tokenization procedure when a vocabulary is obtained or modified. Time-aligned phonetic transcriptions were not performed manually due to the tremendous effort required. Instead, they were derived using a forced alignment procedure during the training process.

2.1.2 Data Analysis

We have collected about 3,100 spontaneous utterances from 64 speakers from different regions of China, 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. The utterances are recorded through telephone channel and sampled at 8 kHz sampling rate.

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 2-1.

	TRAIN	**DEV**	**TEST**
No. of utterances	6,457	500	274
Type of utterances	Spontaneous and read		Spontaneous
No. of speakers	93		6
Words per utterance	8.3	8.5	8.0

Table 2-1: Summary of the corpus

2.2 The Summit Speech Recognition System

The recognition system is configured from the SUMMIT probabilistic, segment-based, speaker-independent, continuous-speech recognition system [13, 36, 37]. SUMMIT explicitly detects acoustic boundaries in order to extract features in relation to specific acoustic segments. The system involves signal processing, segmentation, feature extraction, acoustic modeling, and search components, as shown in Figure 2-1. Language-specific parameters are extracted from training data and are characterized in the form of segment models, lexical models and language models external to the system components.

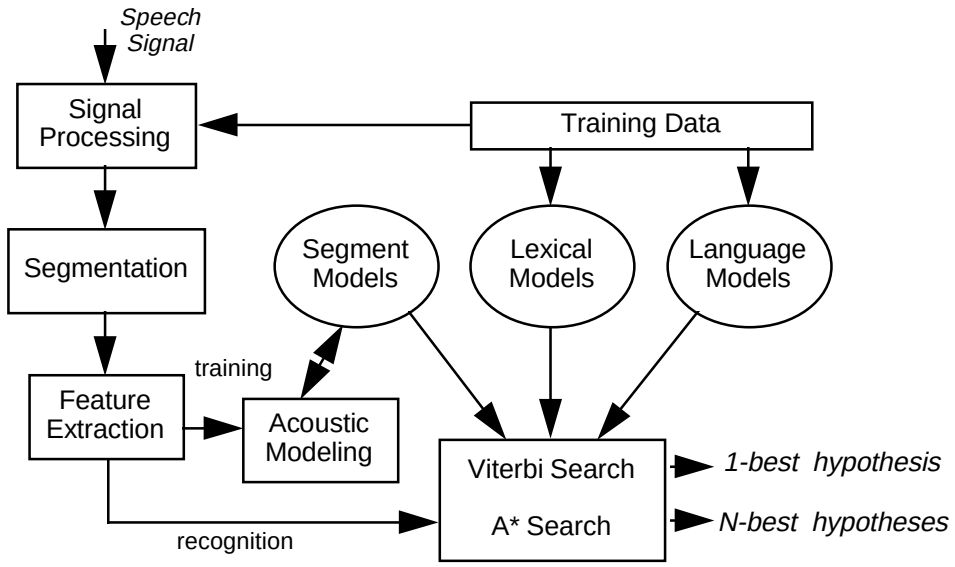


Figure 2-1: Block diagram of the SUMMIT speech recognizer

2.2.1 Signal Processing

The recognition system uses the Mel-Frequency Cepstral Coefficients (MFCC) for signal representation. A set of MFCC coefficients are computed for every 5 ms frame of input speech waveform. Details about the signal processing can be found in [6].

2.2.2 Segmentation

During training, the segment boundaries are provided by the time-aligned phonetic transcription derived from a forced alignment procedure. In recognition, a segmentation algorithm [12] is used to provide a multi-level segment hypotheses associated with scores. Figure 2-2 shows the forced alignment and segment networks for an example Chinese utterance.

2.2.3 Feature Extraction

40 features are extracted for each segment. We use the same feature vector as used by the English recognition system developed in our group. An important reason to choose this feature vector is to maintain consistency, because we need to initialize

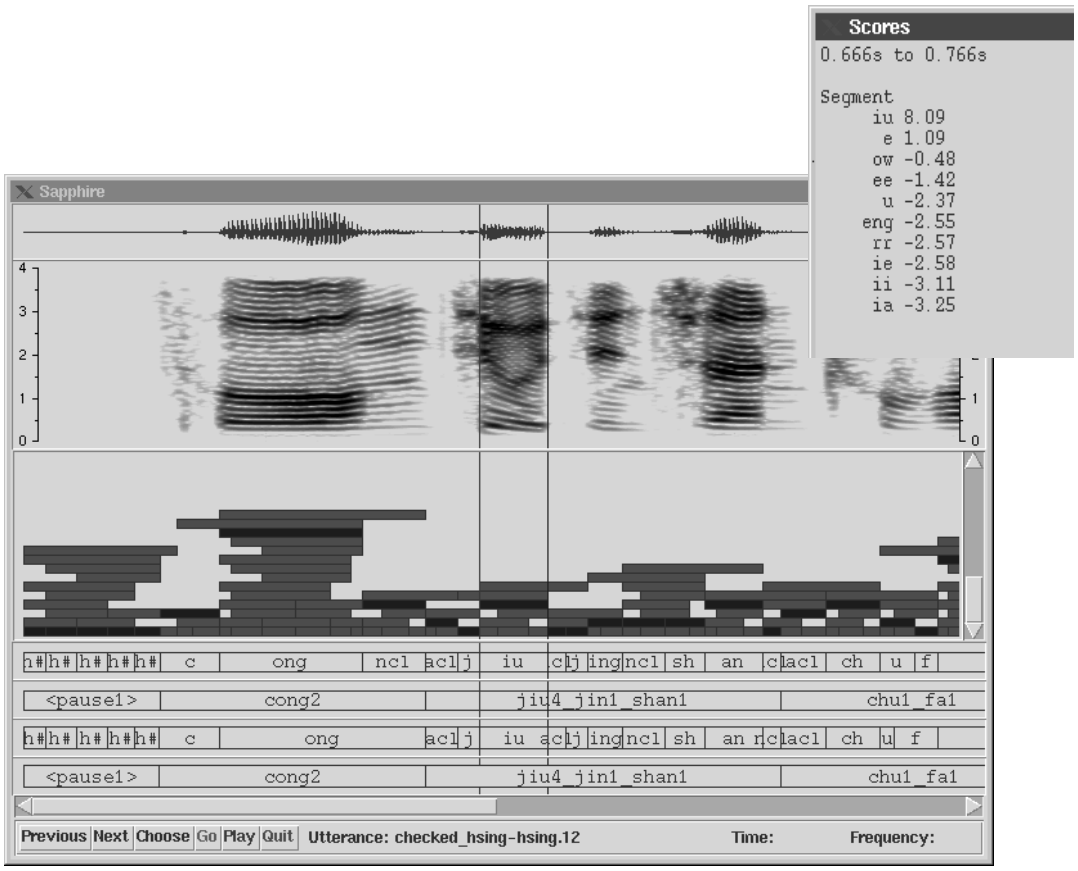


Figure 2-2: The recognition of an example Mandarin utterance. Shown from top to bottom are the waveform, spectrogram, segment hypothesis network (associated with scores), best-paths (both phonetic and orthographic), and forced-paths (both phonetic and orthographic). The small inset window shows the scores of different phone hypotheses for the segment between the two vertical time markers.

Chinese segment models by seeding from English phone models.

2.2.4 Acoustic Modeling

We use mixtures of diagonal Gaussians to model the probability distribution of acoustic features for each of the phones. In recognition, the models are applied to produce phone hypotheses and scores. For example, Figure 2-2 shows the phone hypotheses and scores for the segment between the two vertical time markers. The quality of the models depends on the number of mixtures used, which could be decided empirically depending on the recognition results.

2.2.5 Search

We use a Viterbi search to determine the best path through the network of segment and phone hypotheses and scores, and an A^* search to recover the N-best paths. The N-best hypotheses are further selected by the TINA language understanding system by applying high level syntactic and semantic constraints. A pronunciation network reflecting transition probabilities between phones and a bigram language model reflecting transition probabilities between words are also applied to produce scores during the search.

2.2.6 Iterative Training Process

Since the time-aligned phonetic transcriptions need to be derived through the forced alignment procedure, which requires the existence of a set of models, we have to use the following iterative training procedure. First we derive an initial model for each phone by seeding from an existing similar model. Then we use these models to perform a forced-alignment on the training utterances to get an appropriate alignment in segmentation. Thereafter, we train the phone models using the derived phonetic transcriptions. The last two steps are iterated until convergence in recognition performance. And this process has to be repeated whenever a change in the model set is made.

Monday	xing1 qi1 yi1	li3 bai4 yi1	zhou1 yi1
Tuesday	xing1 qi1 er4	li3 bai4 er4	zhou1 er4
Wednesday	xing1 qi1 san1	li3 bai4 san1	zhou1 san1
...	...	...	...

Table 2-2: Chinese vocabulary for weekday names

2.3 Issues in Mandarin Speech Recognition

The Chinese language has many distinctive characteristics which need to be addressed specifically for speech recognition. We also have a vocabulary mixed with both Chinese and English words, which adds more challenges for acoustic modeling. In this section, we address these issues in more detail.

2.3.1 Vocabulary

It is not always obvious where the word boundaries should be for Chinese; many decisions have to be made based on different considerations. For example, the word "week" can be expressed in three ways in Chinese: "xing1 qi1", "li3 bai4" and "zhou1". And the names for weekdays are formed by attaching a numeric index after the word week (with some variations for the word "Sunday"), i.e., "week one", "week two", etc. for "Monday", "Tuesday", as shown in Table 2-2. So a total of about 22 words are needed to cover the names for all the days in a week. If we separate the numeric index from the week, then the weekday names can be covered by only 10 words, and those 10 words are included in the vocabulary anyway to cover the words for week and digits. In general, we can greatly reduce the vocabulary size by letting words be synthesized from these smaller units. However, this type of representation can not be sufficiently modeled by the class bigram language model which is used in our system. Currently we are omitting the explicit knowledge of these words, recognizing that this is suboptimal in terms of the language model.

We also have English words in our vocabulary due to the application scenarios. The English vocabulary mainly consists of place names and local property names that

either do not have a common Chinese translation, or are preferred over the Chinese translations by some speakers. A multilingual vocabulary is probable in many other applications where information is shared globally, and it is a challenging problem to address for acoustic modeling.

There might exist multiple names for the same semantic entity. For example, "San Francisco" is referred to as "jiu4 jin1 shan1" (meaning "old gold mountain" for historic reasons) in mainland China and "san1 fan2 shi4" (by pronunciation) in Taiwan, but the English name could also be used by some bilingual speakers. These synonyms potentially lead to a large growth in the vocabulary size.

Due to the overwhelming number of English words in the GALAXY knowledge domain, it is not practical to include all of them in the Chinese system. Currently we included English names for some U S cities and states in our vocabulary, because they are likely to be referred to by the user in English. For most of the countries and international capital cities, there usually exist standard Chinese names for them; so the English names are omitted in our vocabulary. As to the place names in the City Guide domain, such as restaurant names, etc., we observed that most users were reluctant to pronounce the "odd" English names; so most of them were also eliminated from our vocabulary. Since there are also no proper Chinese names for them, the user is encouraged to refer to them by index or by clicking.

We derived a vocabulary of about 1,000 words, which is much smaller than the English vocabulary due to the exclusion of many English place names. There are about 780 Chinese words and 220 English words. Each Chinese word has an average of 2.1 characters.

2.3.2 Acoustic Modeling

The acoustic modeling has to deal with both the Chinese and English pronunciations. Since the acoustic units of the two languages are significantly different from each other, it is very difficult to derive an adequate model set to cover both languages without running into sparse training data problems. The lack of hand-labeled time-aligned phonetic transcriptions and good enough initial models for Mandarin speech

also poses challenges to training. We address these problems in more detail in this section.

Model Set

Syllables in Mandarin speech have an initial/final structure, as illustrated in Figure 2-3. There are 21 context-independent initials and 38 context-independent finals in total, and they are usually chosen as the sub-syllable modeling units in Mandarin speech recognition. We believe this model set is particularly suitable for the segment-base SUMMIT system, because the dynamic acoustic properties of the syllable finals can be well captured by the segment feature vector, and they are more robust than the models based on the finer sub-structure of the finals.

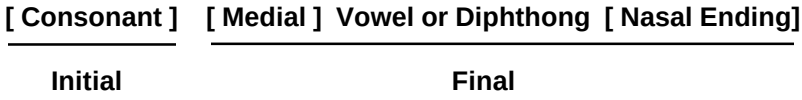


Figure 2-3: Decomposition of syllable structure in Mandarin Chinese

However, we must choose an acoustic model set to cover both the Chinese and English pronunciations. But we can not simply choose our model set to be a union of the Chinese initials and finals, and the English phonemes, because there are many similar subsets and the number of training tokens is inadequate for most of the English phonemes. Our solution is to base our model set on Chinese initial/final inventory, and then add certain English specific models depending on the availability of training tokens. For those discarded models, we have to use existing models or their combinations to approximate them in the pronunciation network. In practice, the phoneme set is adjusted iteratively during the training process, based on performance in segmentation and recognition. Refer to Table 2-3 for our final choice of model set with example occurrences.

There are 23 initials/consonants and 30 finals/vowels in our model set, which are used to specify the pronunciations. An anti-phone unit is used to model all possible forms of non-phonetic segments; refer to [13] for a more detailed description. The

SYMBOL	EXAMPLE	SYMBOL	EXAMPLE
a	b*a*1	ay	b*ai*3
aw	b*ao*4	e	d*e*2
ee	w*e*stern	ey	b*ei*3
ar	*er*4	i	c*i*ty
ia	j*ia*1	ie	j*ie*1
ii	b*i*3	iu	j*iu*4
rr	sh*i*4	o	b*o*1
ow	d*ou*1	u	b*u*4
wa	h*ua*1	wey	h*ui*2
uo	d*uo*1	uu	x*u*4
ue	x*ue*2	an	b*an*1
ang	b*ang*1	eng	b*en*3
ian	b*ian*1	ing	b*ing*1
ong	d*ong*1	wan	d*uan*3
un	d*un*4	uan	y*uan*4
m	*m*a2	n	*n*a4
b	*b*a1	d	*d*a2
p	*p*o1	t	*t*e4
g	*g*en1	k	*k*en3
c	*c*i2	ch	*ch*i1
s	*s*i1	sh	*sh*i4
z	*z*i1	zh	*zh*i1
j	*j*i3	q	*q*i2
x	*x*i3	l	*l*i2
f	*f*u4	v	a*v*enue
h	*h*a1	acl	alveolar closure
lcl	labial closure	vcl	velar closure
ncl	nasal ending	not	anti-phone
iwt	inter- and intra- word silence		
h#	utterance initial and final silence		

Table 2-3: Acoustic-phonetic symbols and example occurrences

remaining 6 units are used to model various types of silence. /acl/, /lcl/, /vcl/ are used to model the silence preceding a stop consonant, i.e., /acl/ for alveolar closure, /lcl/ for labial closure and /vcl/ for velar closure. /h#/ is for utterance initial and final silence, and /iwt/ is for inter- or intra- word silence. /ncl/ is an optional silence following a nasal phone, which was introduced to obtain better segment alignment during the unsupervised training process.

Training process

We have to derive a set of initial models to perform the forced-alignment procedure. Since we do not have any Chinese acoustic models to deploy at the beginning, the phone models have to be seeded from English models. We try to seed a Chinese phone model on its English counterpart if there exists a similar one, and for many of the syllable finals which do not have any close English analog, we seed them from schwa because of its high variability. This method appears to be adequate for the purpose of getting an appropriate initial alignment in segmentation. The acoustic models are then further trained on the acoustic data collected from native speakers of Mandarin Chinese.

Whenever there is a change in the model set, newly added phones need to be seeded and the iterative training process is repeated. One experience we learned about seeding is that it is not always advantageous to seed a new model from a partially similar analogy. For example, when we seed the Chinese syllable /ing/ from /i/, the segmentation in the forced alignment always tries to take only the /i/ part for /ing/ segment, and throw the /ng/ part away as /iwt/ (inter-word silence), or as part of the following segment. Thus the model will not represent the desired acoustic unit after convergence. Instead, we seed them on models with similar substructure(for example both are formed as a diphthong followed by a nasal, etc.), along with an artificial reward for the poorly modeled phonemes and high penalties for /iwt/ insertions in the early stages of iteration. Thus the new models are more likely to occupy the entire span of the target segment during forced alignment, rather than throwing away part of it to an inserted model or the following segment, and the models will converge to

the right segment after several iterations.

Choosing the right models and training them properly proved very critical to bringing up the recognition performance. The word error rate was reduced by about 10% on development after many rounds of changes and retraining of the phonetic models.

Mixture Number

We use mixtures of diagonal Gaussians to model the probability distribution of the acoustic features for each phone. An experiment was performed to decide the best mixture numbers depending on recognition performance and speed. We vary the maximum number of mixtures for all the models and observe the change in word error rate on the development data set. Figure 2-4 shows that the word error rate first drops with more mixtures in the models, but rises slightly with further increasing. The reason is that the models become more and more accurate with increasing of parameter number. However, the models might be over-fitted to the training data with too many parameters. They start to capture the noise components in the training data, and may not be general enough for unseen data. We finally choose the maximum number of mixtures to be 60 because it gives us the best performance on development data and the recognizer still runs in roughly real time.

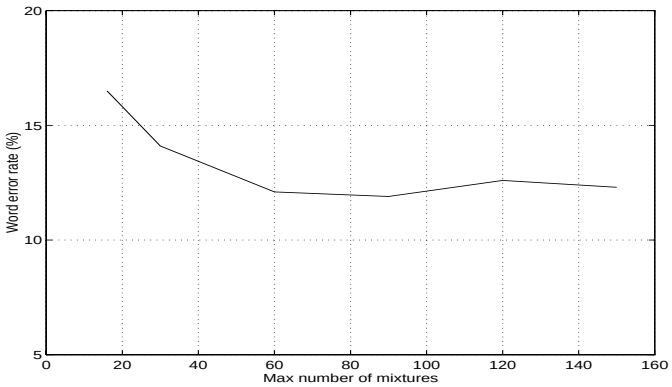


Figure 2-4: Recognition performance with varying maximum number of mixtures

2.3.3 Dialectal Variations

There are many very diversified dialects in China. A dialect in one place could be totally a "foreign language" to people from another, and the phoneme set of a dialect also varies from one to another. Although Mandarin is the standard official language, and each word usually has only one pronunciation, there are many possible variations on the sub-syllable level due to dialectal influences. In this section we describe some of the prominent variation patterns, and discuss two approaches to handling them.

Analysis of Variation Patterns

Some variations occur at syllable initials. For example, there are several non-retroflex/ retroflex fricative pairs in the Mandarin initial set, such as "z/zh", "c/ch", and "s/sh". People from the northern part are usually able to distinguish them, but many people from the south have difficulty making the distinction, because the retroflexed sounds do not exist in their dialects. The same is true for the "l/n" pair; many dialects in southwestern China do not have the /n/ sound as initials. There are two ways that people can make mistakes for each pair. We do not have statistics of these errors in our corpus, but it is observed that non-retroflex fricatives substituting for retroflex fricatives and /n/ substituting for /l/ occur more often than the other way around.

Variations can also occur at syllable finals. For example, people from the Beijing area tend to retroflex certain finals, sometimes at the end of a word, and sometimes in the middle of a word. It is very hard to characterize the rules for this type of variation, because they usually depend on both the acoustic and the linguistic environment. Table 2-4 shows some examples.

Handling Variations in the Recognizer

Variations can be modeled at several levels in the recognizer. One way is to enumerate all the possible pronunciations for a word in the lexicon (or equivalently use phonological rules to summarize the variation patterns). For example,

cheng2_shi4	(ch , c) eng (sh , s) rr

Chinese	**English**	**Standard**	**Variant**
cheng2_shi4	city	*ch* eng *sh* rr	*c* eng *s* rr
zhi1_dao4	know	*zh* rr d aw	*z* rr d aw
na3_er2	where	*n* a ar	*l* a ar
fan1_guan3	restaurant	f an g *wan*	f an g *wa er*
hua1_dian4	florist	h *wa* d ian	h *wa er* d ian

Table 2-4: Examples of dialectal variations

zhi2_da2	(zh , z) rr d a
ya4_te4_lan2_da4	y ia t e (l , n) an d a
fan4_guan3	f an g (wan , wa er)

Although this approach is very straightforward, it was not successful in our system for the following reasons. Because we do not have hand-aligned phonetic transcription in our corpus, the recognizer has to perform a forced-alignment to judge which path the pronunciation should take for each word and where the segment boundaries are. However the initial models are not good enough to guarantee correct classification in forced-alignment. In fact, the classifier often makes the same errors in choosing alternatives in the pronunciation. For example, /n/ is often classified as /l/ and vice versa. As a result, the training tokens for /n/ and /l/ are mixed with both /n/ and /l/ realizations, and the models are also a random mixture of both real /n/ and /l/.

The second approach is to provide only the standard pronunciation in the lexicon and let the acoustic model capture the variations. We justify this approach for the following reasons:

1. We use mixtures of diagonal Gaussians for acoustic modeling, which are capable of modeling mixed probability densities.

2. The correct tokens are dominant for each model. For example, even though there are both /s/ and /sh/ tokens for the model /s/ and the model /sh/, the /s/ tokens are dominant for the /s/ model and the /sh/ tokens dominant for

the /sh/ model. So it is reasonable to expect the models for /s/ and /sh/ to be significantly different and dominated by the "correct" components.

3. Though the models are still mixed, it is better than the near random mixing produced via the unsupervised training process.

We observed a 1.5% reduction in word error rate on the development data set after we disabled the alternatives of "z/zh", "c/ch" and "s/sh" in the lexicon and retrained the models.

2.3.4 Tokenization

Each word in the Chinese language is composed of one or multiple characters. Given a sentence in characters, the process of segmenting character sequences into word sequences is called tokenization. Since the same characters are shared across different words (the chance of sharing is even higher if the representation is in pinyin instead of ideography), the boundaries between words could be ambiguous in many cases, which makes correct tokenization a challenging problem. In this section, we analyze the impact of this problem on our system, and propose an approach to solving this problem.

The problem of tokenization could have been avoided if we could choose syllables as the vocabulary units. However, it is not practical to do so for the following reasons:

1. Syllable recognition is difficult because of the existence of many confusable subsets of syllables.

2. There is a higher chance of homophones at the syllable level than at the word level.

3. Words are semantic units, and language modeling is more natural at the word level than at the syllable level.

We divide tokenization errors into two types, as illustrated in Figure 2-5. The type 1 errors are easy to detect, and can actually be prevented during the search

- Type 1, error in tokenization causes out-of-vocabulary words, e.g.

dao4 da2_la1_si1	to Dallas
dao4_da2 la1 si1	arrive ? ?

- Type 2, error in tokenization does not cause out-of-vocabulary words, e.g.

jiu3 ba1	nine eight
jiu3_ba1	bar

Figure 2-5: Summary and examples of tokenization errors

u_a	er4	jiu3	ba1	hao4	hang2_ban1	de5	piao4_jia4	shi4	duo1_shao3
(u_a	two	nine	eight	number	flight	+s	fare	is	how much)

u_a	er4	jiu3_ba1	hao4	hang2_ban1	de5	piao4_jia4	shi4	duo1_shao3
(u_a	two	bar	number	flight	+s	fare	is	how much)

Figure 2-6: Example of different ways of tokenization

stage in recognition. For example, since "la1" and "si1" are not vocabulary items, the second way of tokenization is not a valid path in the pronunciation network and thus is discarded in the search process. However, the type 2 errors are more subtle and need to be detected in a broader context. For example, the recognizer is likely to propose either of the word sequences as shown in Figure 2-6.

Applying our knowledge of the Chinese language we know that only the first sentence makes sense. Although one can hope that the language model will favor the correct choice, the constraints of a bigram are not robust enough. We use the TINA language understanding system for a more reliable solution. For the word sequence proposed by the recognizer, we take out the tokenization information and let the TINA system parse the character sequence instead, thus recreating the tokenization. We believe that the tokenization performed by TINA implicitly is more reliable because more extensive linguistic constraints are used in parsing. This approach also obviates the tedious task of maintaining the consistency between the words defined in the recognizer vocabulary and the words used in the TINA grammar if we were to use words as the terminal nodes in the grammar. We will discuss this issue further in Chapter 3.

2.3.5 Homophones

The Chinese language is a tonal syllabic language, in which each character is pronounced as a mono-syllable with a particular tone. Considering that there are about 6,000 commonly used characters and only about 1,300 tone-syllable pairs, it is quite usual that several distinct characters may map to the same tonal syllable, forming homophones. When we only utilize the 416 base-syllables to represent the pronunciation, then the chance of homophones is even higher. For words that contain multiple characters, homophones rarely occur at the word level.

Analysis of Homophones

We adopt the approach of performing only the base syllable recognition, mainly because the tone recognition function is not implemented in the SUMMIT system which was initially designed for English recognition. We also suspect that ignoring tone recognition will not influence the system performance very much for the following reasons:

1. Homophones exist even with perfect tone recognition.

2. There are only 18 additional homophone pairs when we ignore tone recognition.

3. Homophones caused by ignoring tone recognition can be resolved using the same method for true homophones, i.e., by relying on higher level linguistic knowledge.

The homophones can be divided into two types. One type involves words that are the same in tone-syllable pairs. Those words are merged as one identity in the lexicon because of our tonal-pinyin representation. For example, both the noun "city" and the verb "be" correspond to "shi4" in the lexicon, and are thus totally indistinguishable to the recognizer. We will show that this type of ambiguity can be easily resolved by the language understanding system. The other type of homophone is caused by ignoring tone recognition. This type involves words that are the same in base-syllables but differ in tones. Even though those words are distinctive entries in the lexicon,

qing3	gao4_su4	wo3	xi1_ya3_tu2	*shi4*	de5	tian1_qi4	yu4_bao4
(Please	tell	me	Seattle	city/be	+s	weather	forecast)

Figure 2-7: Example of type 1 homophones

Cambridge	*li2*	you3	bo2_wu4_guan3	ma5
(Cambridge	in	have	museum	[question])
Cambridge	*li3*	you3	bo2_wu4_guan3	ma5
(Cambridge	from	have	museum	[question])

Figure 2-8: Example of type 2 homophones

they have the same pronunciation. For example, "li2" (from) and "li3" (in) are both pronounced as [l ii]. The second type of homophone has more constraints than the first type because they are distinctive in the bigram language models. However, relying on the bigram to distinguish them is not robust. We will show that this type of homophones can be resolved in a similar manner as the first type.

Resolving Homophones Using TINA

We rely on the TINA language understanding component to solve both types of homophone problems. The solution to the first type is actually embedded in the system. For example, when the recognizer proposes the word sequence shown in Figure 2-7, the recognizer does not know whether "shi4" is the word "city" or something else. But when the sequence is parsed by TINA, the meaning "city" is naturally chosen given the surrounding context in the sentence.

The solution to the second type of homophones is based on the same approach, but needs some extra processing. For example, it is highly probable for the recognizer to propose the top sentence in Figure 2-8, while the correct choice is the bottom one. We propose the algorithm shown in Figure 2-9 to correct this type of error.

The effectiveness of this algorithm depends on the assumption that one and only one of the alternatives generates a successful parse. We observe that out of the 5.3% substitution error rate (273 occurrences) in the development set, less than 10 occurrences are caused by tonal substitutions. So the type 2 homophone problem is

- Summarize type 2 homophones in our vocabulary to form confusable subsets.
- If a sentence that failed to parse contains words in the above sets, substitute in the alternatives until success.
- If no alternatives left, correction fails.

Figure 2-9: Algorithm to correct type 2 homophones

not a significant error source. Actually there are many other single syllable words in our vocabulary forming confusable subsets. We will describe a post-processing method based on the TINA language understanding system in Chapter 3.

2.4 Summary

In this chapter, the basic configuration of the recognizer was briefly described and some of the problems particular to Chinese were discussed in more detail. The phoneme set was formed to cover both Chinese and English acoustic models in our vocabulary. Dialectal variations in pronunciation were incorporated into the phone models. We rely on the language understanding system TINA for a robust processing of tokenization and homophones. A formal evaluation of the recognition on unseen test data will be reported in Chapter 5.

Chapter 3

Natural Language Understanding

Natural language understanding for Chinese is processed using the TINA system [29]. In TINA, an initial context-free grammar [1] written by hand is first converted to a network structure, with transition probabilities trained from a set of example sentences. The parser uses a stack decoding search strategy, with a top-down control flow. A feature-passing mechanism is included to deal with long-distance movement, agreement, and semantic constraints. TINA uses the same parsing mechanisms for all languages, while the language specific parameters are specified in grammar rules, augmented with a set of features used to enforce syntactic and semantic constraints. The grammar rules interleave semantic categories and syntactic categories. We believe that the approach of combining syntactic and semantic knowledge in the grammar is very suitable for parsing Chinese, because there are few inflectional or grammatical markers in the language, which makes purely syntactic approaches almost impossible [7].

In this chapter, we first introduce the method of porting grammar rules from English to Chinese for the GALAXY domain. There are some characteristics of the Chinese language that challenge the multilingual approach of the TINA parser. We analyze some of the difficulties and propose our solutions. Next, some of the special processing mechanisms in TINA applied for Chinese are described. After that, performance on the spontaneous training data is measured in terms of the parsing coverage and the quality of the semantic frame. Finally, the interface between the

recognizer and the parser is described, and some experimental results on development data analyzed. Evaluation on unseen test data will be reported in Chapter 5.

3.1 Grammar Rules

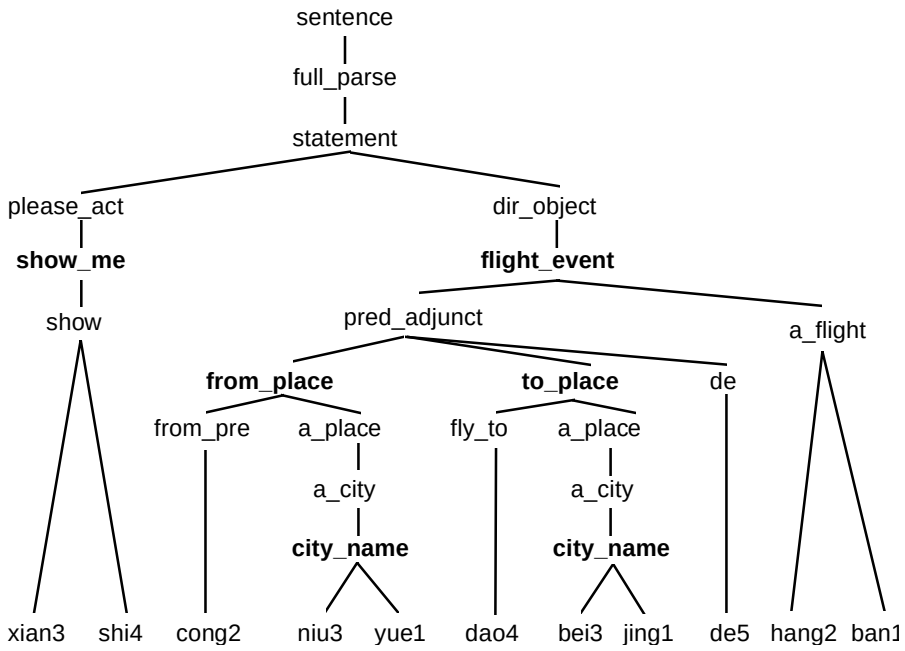
The TINA grammar rules are written such that they describe both the syntactic and semantic structures of the language. The syntactics are embedded in the structure of a parse tree; and the semantics of the sentence are completely encoded in the names of a subset of the categories in the parse tree; thus no separate semantic rules are needed to perform further annotation. The names and the hierarchical relationship of the meaning-carrying categories are then used to extract a semantic frame as the meaning representation of the sentence. Figure 3-1 shows the parse tree and the corresponding semantic frame for an example Chinese sentence, with the meaning-carrying categories emphasized in bold in the parse tree.

3.1.1 Writing Grammar Rules

The grammar is built from a set of training sentences gradually. Initially, each sentence is translated by hand into a list of the rules invoked to parse it. After the grammar has accumulated a substantial knowledge of the language, many new sentences can be parsed automatically, or with minimal increment of new rules.

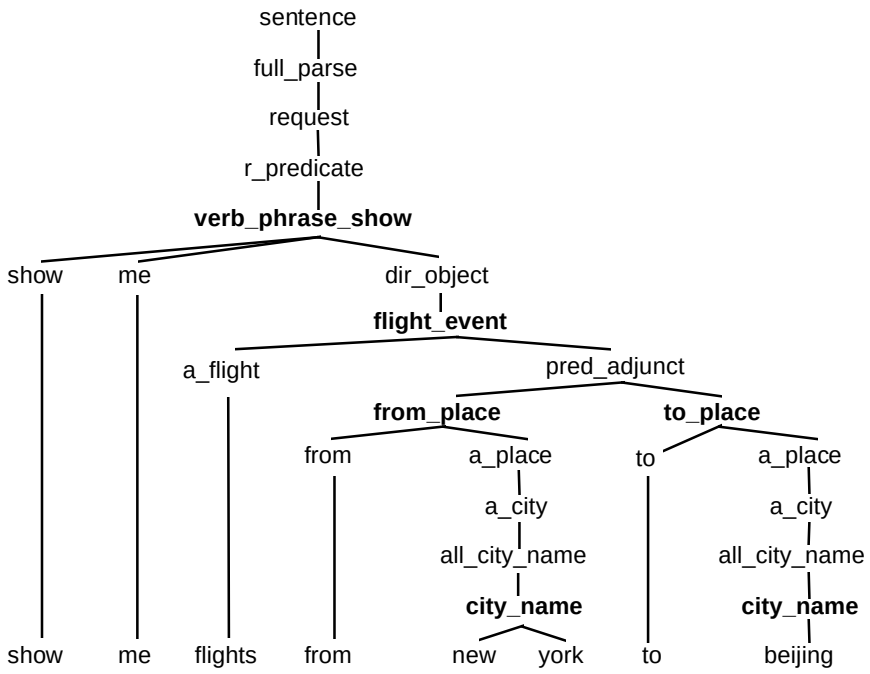
In writing the grammar rules for Chinese, the following strategy is adopted to take advantage of the existing English grammar. The closest English translation of the Chinese sentence is first parsed, using the English grammar, to generate a parse tree and semantic frame. Then the Chinese grammar is written in a way that maintains the internal nodes of the parse tree as similar to those for English as possible while respecting the different word order and sentence structure of Chinese. The resulting semantic frame should be the same as that of the English counterpart. This method has proven to be very efficient and consistent for both syntactic and semantic analysis. Figure 3-1 and Figure 3-2 illustrate this method with examples of parse trees and semantic frames for a Chinese sentence and its English counterpart. Notice

that the set of meaning-carrying categories are very similar in the two parse trees, even though the parse trees themselves look quite different from each other; and the semantic frames are almost the same, except for the "singular/plural" information, which is usually missing in Chinese. The language dependent syntactic information is inconsequential to the understanding task; they are included in the semantic frame mainly to enable the language generation system to generate a correct paraphrase for the input sentence.




```
{c display
    :topic {q flight
                :pred {p from
                            :topic {q city
                                        :name "new york" } }
                :pred {p to
                            :topic {q city
                                        :name "beijing" } } }
    :domain "AirTravel" }
```

Figure 3-1: Parse tree and semantic frame for the Chinese sentence "Show from New York to Beijing +s flight."




```
{c display
    :topic {q flight
                :number "pl"
                :pred {p from
                            :topic {q city
                                        :name "new york" } }
                :pred {p to
                            :topic {q city
                                        :name "beijing" } } }
    :domain "AirTravel" }
```

Figure 3-2: Parse tree and semantic frame for the corresponding English sentence "Show me flights from New York to Beijing."

3.1.2 Special Issues in Parsing Chinese

The TINA system tries to use the same parsing mechanism for different languages while relying on the grammar rules to handle the language specific information. However, the Chinese language has some characteristics that are substantially different from English, which could be problematic for this approach. For example, Chinese words are composed of strings of characters without spaces marking word boundaries. Although TINA is capable of extracting the meaning of the sentence without explicitly performing word segmentation, inefficiency is induced because of the ambiguity in word boundaries. The problem of inefficiency is exacerbated by the ambiguity of homophones, because the homophones are indistinguishable in pinyin representation. In addition, the Chinese language is left-recursive. While the top-down control flow of TINA is suitable for a right-recursive language like English, it has some inherent difficulties with a left recursive grammar. We address these issues in more detail in the following.

Ambiguity of Word Boundaries and Homophones

We mentioned in Chapter 2 that we decided to rely on the grammar rules of TINA for more robust tokenization and homophone disambiguation. There are also several other practical considerations. One important reason to choose pinyin to specify the grammar is that Chinese ideography has a different binary representation (16 bits) in the computer, which would require rewriting much of the TINA code in order to process it properly. In addition, the terminal nodes of the grammar can be specified in pinyin syllables, thus obviating the tedious task of manually maintaining the consistency of "words" specified in recognition vocabulary and the grammar rules.

Since the TINA grammar is heavily constrained by semantic categories throughout the parse tree, it is usually able to reconstruct the correct tokenization of the sentence and resolve homophones. For example, the two syllables "jiu3 ba1" in the sentence shown in Figure 2-6 could be interpreted as either two individual words "jiu3" (nine) and "ba1" (eight) or a single word "jiu3_ba1" (bar) without considering the context

in the sentence, due to the ambiguities in both word boundary and homophones. However, when it is captured under the semantic category "flight_number", as shown in the parse tree in Figure 3-3, the latter possibility is naturally ruled out by the semantic constraints.

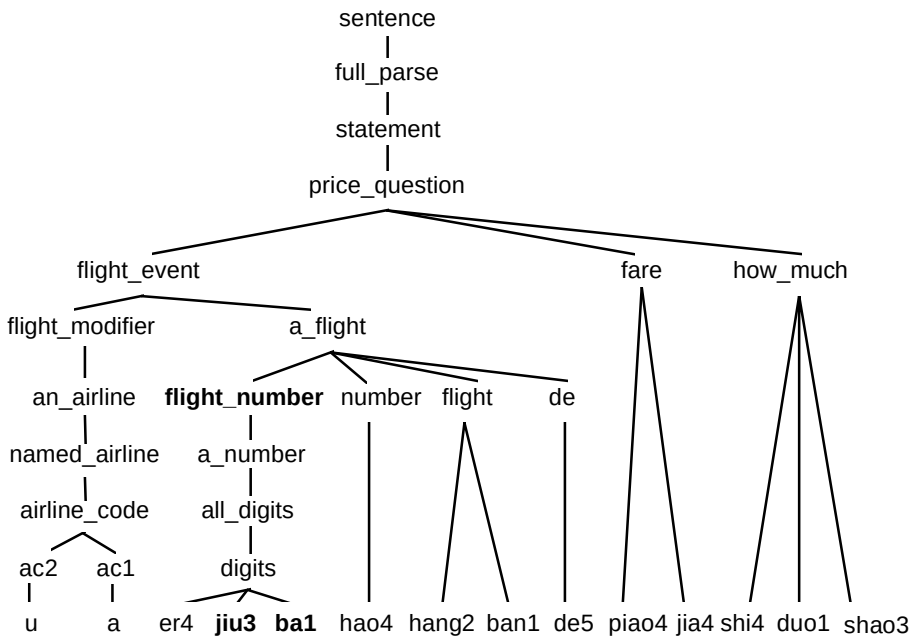


Figure 3-3: Example parse tree illustrating tokenization and homophone disambiguation

However, parsing syllable sequences has the disadvantage of causing inefficiency. This is because the Chinese words are formed by combining characters, and many different words could start with the same character. The extent of sharing is even higher with the pinyin representation, because homophones, which are different in ideography, become indistinguishable in pinyin form. For example, "jiu3 ba1" (bar), "jiu3 dian4" (liquor store), and a digital string beginning with "jiu3" (nine) all start with "jiu3", the "jiu3" in the first two words corresponds to the same ideography, but is different from the one in the third word. Given the sentence "*jiu3 ba1* san1 de5 di4 zhi3 shi4 she2 me5?" (the third bar's address is what), the parser has to look up all the theories that could start with "jiu3", regardless of what is following. Since the digits are very popular in the grammar rules, there could be an explosion of active

nodes in the early search stage which may even exceed the current limit of the parser, which could be avoided if the parser can foresee the whole word "jiu3 ba1", or the "jiu3" is specified in ideography such that it is different from the "jiu3" as a digit.

We use the following approach to reduce inefficiency. We perform a partial tokenization on the input pinyin sequence as a preprocessing measure. It is implemented using a rewriting mechanism provided by TINA. If the input sentence contains strings of pinyin that are specified in the rewrite rules, they are replaced by the corresponding "words", usually in the form of syllables connected by hyphens. We apply this processing only to a few words that start with problematic syllables, and special caution is taken to avoid tokenizing strings of characters that could be ambiguous in word boundaries. We can also view this approach as providing a "look-ahead" mechanism by hyphenization, under the limit that only words without ambiguities are processed in this way.

We can also use "indexed pinyin" representation to reduce ambiguity at word initial position caused by homophones. For example, homophones of "jiu3" can be differentiated by "jiu3*0","jiu3*1", etc., as most of the Chinese input softwares do. However, it is very tedious to maintain the consistency of index in the lexicon and grammar. We did not apply this method in our current system because the parser is able to run in real time with the "partial tokenization" approach.

Since TINA has a probabilistic framework, efficiency can be further improved by optimizing probability assignments in the grammar. We observed about 25% improvement in speed after training the grammar using the collected data and some artificial sentences. The parser uses less than 0.07 second on average to process a sentence after training. [1]

Left-recursion

Chinese is a left-recursive language, with nouns preceding their prepositions in the prepositional phrases, and prepositional phrases preceding the noun they modify.

[1] Tested on a Pentium Pro 200 PC in batch mode.

This potentially leads to a valid Chinese phrase like "MIT fu4 jin4 de5 dian4 ying3 yuan4 fu4 jin4 de5 hua1 dian4" (corresponding to the English phrase "the florists near the cinemas near MIT"), and the prepositional phrase can keep going indefinitely in theory. The left part of Figure 3-4 shows the left-recursive grammar for such a Chinese phrase.

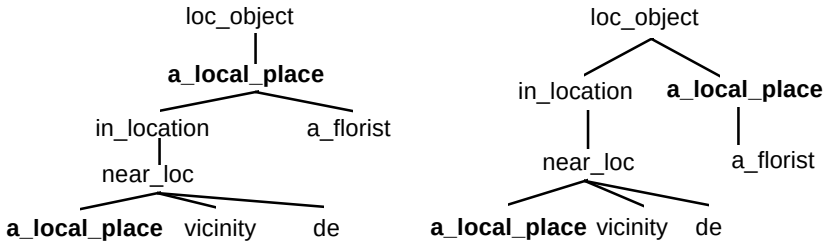


Figure 3-4: A left-recursive grammar and the non-recursive implementation

The left-recursive grammar and the top-down control flow of TINA can potentially lead to infinite recursion. Even though the probability becomes very small as the chain keeps looping, it still wastes a lot of parse space building implausible structure. Our current method to handle this problem is to avoid writing such a grammar. Notice that even though the left-recursion is valid in theory, the loop of prepositional phrase seldom occurs more than once in a normal sentence. So we break down the recursion in the grammar and specifically allow the prepositional phrase to go before a noun only once (or multiple times if necessary), as shown in the right side of Figure 3-4. In order to prevent writing left-recursive grammar rules unintentionally, a new feature has been added to TINA to detect such a grammar and give a warning message when the nesting is deep. We have not observed parsing failure on training data due to the absence of a left-recursive grammar.

3.2 Special Processing

TINA also provides various functions in addition to general parsing mechanisms to facilitate special processing. Most of them can be directly applied to Chinese, while

a few need some modification because of the language dependent nature. In this section, we describe some of the functions in detail with their adaptation to process Chinese.

3.2.1 Long-distance Movement

Long-distance movement is very common in English wh-questions, as illustrated in the sentence "(*which street*)$_i$ is the Hyatt on (t_i)?". Unlike English, there is no movement involved in Chinese questions. The questions are usually in the same order as the corresponding statements, with a wh-word replacing the component that the question is about. Figure 3-5 shows some examples of questions and corresponding statements in Chinese.

Question:	Hyatt	zai4	na3 tiao2	jie1	shang4		
	Hyatt	at	what [particle]	street	on		
Statement:	Hyatt	zai4	main	jie1	shang4		
	Hyatt	at	main	street	on		
Question:	san1	jiu3	ba1	hao4	hang2 ban1	ji3 dian3	qi3 fei1
	three	nine	eight	number	flight	when	depart
Statement:	san1	jiu3	ba1	hao4	hang2 ban1	san1 dian3	qi3 fei1
	three	nine	eight	number	flight	3 o'clock	depart
Question:	ni3	da3 suan4	ji3 dian3	cong2	na3 er2	chu1 fa1	
	you	want to	when	from	where	leave	
Statement:	wo3	da3 suan4	san1 dian3	cong2	bo1 shi4 dun4	chu1 fa1	
	I	want to	3 o'clock	from	Boston	leave	

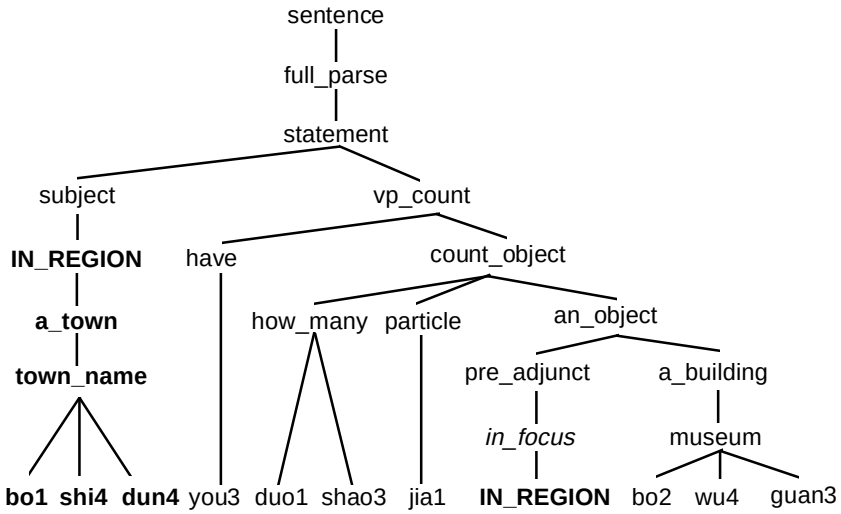
Figure 3-5: Examples of Chinese sentences without long-distance movement

However, long-distance movement is very common in the "there are ..." type of sentences in Chinese, no matter in statement form or question form, in which various adverbial phrases can be moved to the beginning of the sentence in a process referred to here as "topicalization," usually with the preposition dropped. Figure 3-6 shows some examples of such sentences. Notice that the question and the corresponding statement still have the same word order.

The long-distance movement phenomenon in Chinese can be solved using the same mechanism as for English wh-questions. We use the first sentence in Figure 3-6 as an

Question:	(*bo1 shi4 dun4*)$_i$	you3	duo1 shao3	(t_i)	bo2 wu4 guan3		
	Boston	have	how many		museum		
Statement:	(*bo1 shi4 dun4*)$_i$	you3	liu4 jia1	(t_i)	bo2 wu4 guan3		
	Boston	have	6 [particle]		museum		
Question:	(*M I T fu4 jin4*)$_i$	you3 mei2 you3		(t_i)	zhong1	can1 guan3	
	M I T near	have not have			Chinese	restaurant	
Statement:	(*M I T fu4 jin4*)$_i$	you3		(t_i)	zhong1	can1 guan3	
	M I T near	have			Chinese	restaurant	
Question:	(*san1 yue4 shi2 wu3 hao4*)$_i$	you3	na3 xie1	(t_i)	hang2 ban1		
	March fifteen date	have	what		flight		
Statement:	(*san1 yue4 shi2 wu3 hao4*)$_i$	you3	zhe4 xie1	(t_i)	hang2 ban1		
	March fifteen date	have	these		flight		

Figure 3-6: Examples of Chinese sentences with long-distance movement




```
{c count
    :topic {q museum
                :particle "jia"
                :pred {p in
                            :topic {q town
                                        :quantifier FOCUS
                                        :name "Boston" } } }
    :domain "CityGuide" }
```

Figure 3-7: Parse tree and semantic frame for an example Chinese sentence with long-distance movement

example to help explain how it works. The parser makes use of two special registers, passed from node to node, the *current focus* and *float-object*. Referring to its parse tree shown in Figure 3-7, the [subject] "bo1 shi4 dun4" generates its subparse as the current-focus. The [vp_count] is an *activator*; it moves the current-focus to the float-object slot, which, when occupied, means there is a gap somewhere in the future to be filled by the partial parse tree occupying the float-object slot. The parse succeeds only when the float-object is taken up by an *absorber*, in this case the [in_focus] modifier for [a_building] (Refer to [29] for more details).

There are two additional points in this parse tree worthy of attention. One is that even though "bo1 shi4 dun4"(Boston) appears to be a noun in the sentence, it should be viewed as a prepositional phrase meaning "zai4 bo1 shi4 dun4"(in Boston). This underlying semantic meaning of "in Boston" is captured by the category [in_region]. Another observation is that [in_region] could be absorbed by the node [pre_adjunct]. However, an extra layer [in_focus] is used to tag a special mark "FOCUS" to "Boston" in the semantic frame, so that the language generation system can move the "Boston" to its proper place when generating a paraphrase for the sentence. We will return to this issue in more detail in Chapter 4.

3.2.2 Keyword Mapping

Keyword mapping or translation is used to map various possible expressions for an identity to the unique keyword in the semantic frame. In the English system, it is mainly used to map nicknames to the standard expression. For example, in the English sentence "Show me the weather for Philly", this mechanism is used to rename the city name "Philly" as "Philadelphia" in the semantic frame.

When parsing Chinese, although the semantics of the sentence are captured in the internal nodes of the parse tree in a language independent form, the keywords are entered from terminal nodes which are in the specific language. In order to maintain the language-independent feature of the semantic frame and backend system, it is necessary that the semantic frame resulting from parsing a Chinese sentence should not contain any Chinese specific expressions. Since the semantic frame is English

based, the translation mechanism is used more extensively in the Chinese system. For example, there are many keywords the backend system uses, such as property names, place names, date names, cuisine types, etc, and they all need to be translated from Chinese into English in the semantic frame. The parsing example shown in Figure 3-1 would result in the semantic frame as shown in Figure 3-8 without the translation mechanism, in which the keywords for city name "New York" and "Beijing" are in Chinese pinyin.

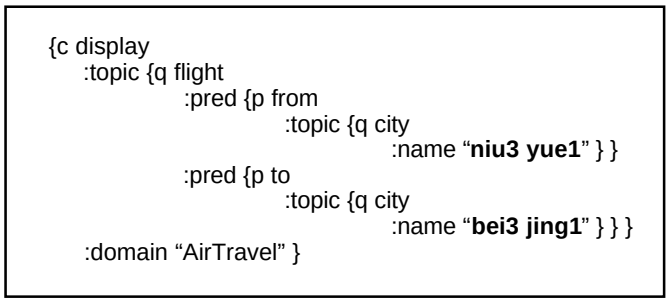


Figure 3-8: An example semantic frame without keyword mapping

When there are a large set of keywords that need to be translated into English, it is highly probable that the same Chinese word has multiple translations in different environments. For example, the two sentences in Figure 3-9 generate the semantic frames as shown in Figure 3-10 before translation. "zhong1 guo2" should be translated as "China" in the left frame, and "Chinese" in the right one. The keyword mapping mechanism can achieve this environment-sensitive translation by utilizing the contextual information already obtained in the semantic frames. "zhong1 guo2" is translated as "Chinese" under the category of "cuisine" and "China" under the category of "country", resulting in the semantic frames in Figure 3-11.

xian3 shi4	bo1 shi4 dun4	de5	zhong1 guo2	can1 guan3
display	Boston	+s	Chinese	restaurant
ming2 tian1	zhong1 guo2	de5	tian1 qi4	zen3 me5 yang4
tomorrow	China	+s	weather	is what

Figure 3-9: Example sentences to illustrate context-dependent keyword translation

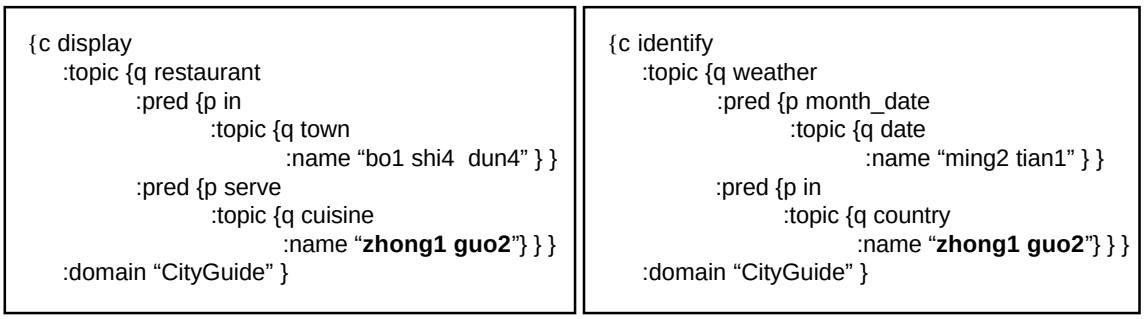


Figure 3-10: Examples of semantic frames without keyword translation

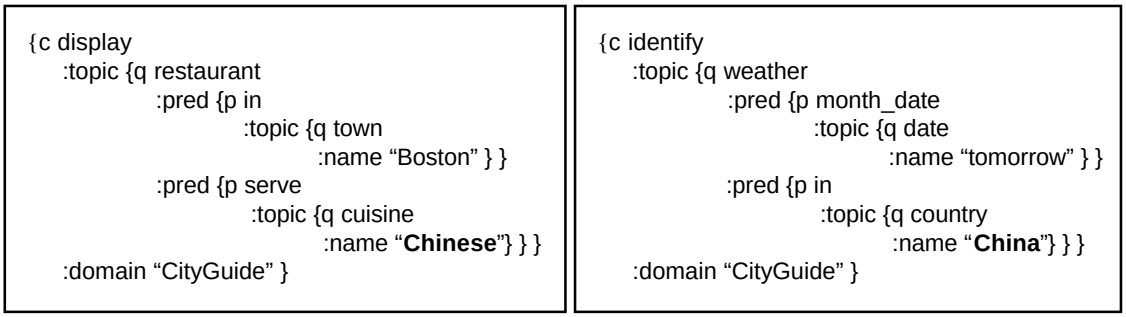


Figure 3-11: Examples of semantic frames with keyword translation

3.2.3 City-state Decoding Algorithm

The city-state decoding algorithm is an important function of TINA to help constrain the recognizer outputs. It would be of great benefit if TINA could know explicitly which cities belong in which countries/states. Thus by avoiding over-generalization of “city state” or “city country” combinations directly in the grammar rules, TINA can provide a significantly more constrained language model for the recognizer. For example, “Orlando Florida” and “Beijing China” are legitimate city names, while “Orlando New York” or “Beijing England” do not make sense in our application domain. If we allow the following grammar rules,

place_name $\Rightarrow$ city_name

place_name $\Rightarrow$ city_name state_name

place_name $\Rightarrow$ city_name country_name

there will be tremendous over-generalization by illegal cross combinations. So we list all the legitimate combinations explicitly as complete city names instead, and rely on a specialized "pick-apart" module to separate out country name or state name if there is one. The pick-apart is also used to aid in discourse resolution. For example, when the user asks for weather information for London, it is not clear whether the user refers to the London in England or Canada. But if the user mentioned a country name before, the pick-apart can be used to decide if the country name should be inherited by the current city name depending on the legitimacy of the combination. If the discourse has country name "England" in its memory, we can safely assume "London" means "London England". But if the discourse remembers "China" as the most recently mentioned country, we need further clarification from the user to decide which "London" the user is referring to.

Two problems occurred when applying this function directly in processing Chinese. One is that a place name in Chinese normally goes in the order of (country state city) as opposed to (city state country) in English, though it is also possible for a user to say (city state) when referring to a US city because of the influence of English customs. So the pick-apart function was enhanced to be able to process both ways. The second problem is associated with the characteristics of Chinese translations of foreign place names. Usually the Chinese translations of foreign names are very long and contain many common characters. For example, "Brasilia" is "ba1 xi1 li4 ya4", and "*Budapest* Hungary" is "xiong1 ya2 li4 *bu4 da2 pei4 si1*". It is usually sufficient to check only the first and last word in the English city name to separate the city and country/state, however, it is inadequate for Chinese because of the character representation of terminal nodes. Two approaches were used to solve the problem. First, more extensive judgments are added in the pick-apart function to make it more rigorous. Second, problematic place names are tokenized using the same approach introduced in Section 3.1.2 to reduce ambiguity. This approach is very suitable for long words because it is not likely to make tokenization errors when there is a match of a long sequence of characters.

3.3 Performance Analysis

We use parsing coverage on all the orthographies of the spontaneous utterances in the training set and development set to examine the performance of the parser. The sentences are divided into the following three categories.

Out-of-domain

The "out-of-domain" category contains queries that exceed the capability of the back-end system. Some of them are beyond the application domains of the system, such as the first sentence in Figure 3-12. Some are semantically wrong, for example, the second sentence in Figure 3-12. Some contain details that can not be processed by the backend system even with perfect understanding. For example, the system does not know about Northeastern University, so the third sentence in Figure 3-12 can not be processed even if a perfect semantic frame is generated.

1.	Charles he2	shang4	you3	ji3	sou1	you2 ting3
	Charles river	on	have	how many	[particle]	yacht
	How many yachts are there in Charles river?					
2.	MIT	you3	na3 ji3 ge4	xi4 yuan4		
	MIT	have	what	cinema		
	What cinemas are there in MIT?					
3.	wo3	xiang3	cong2	dong1 bei3 da4 xue2	chu1 fa1	
	I	want to	from	Northeastern University	leave	
	I want to leave from Northeastern University.					

Figure 3-12: Examples of out-of-domain sentences

Disfluency

The "disfluency" category contains queries that are agrammatical, such as utterances that are truncated or contain stutters. Figure 3-13 shows some examples of disfluent sentences.

1.	qing3	wen4	cong2	bo1 shi4 dun4	dao4	bei3 jing1	you3	...	
	please	ask	from	Boston	to	Beijing	have	...	
2.	qing3	gao4 su4	xia4	yue4	er4 yue4	san1	hao4	de5	tian1 qi4
	please	tell	next	month	February	3	date	+s	weather

Figure 3-13: Examples of disfluent sentences

Clean

The remaining "clean" category contains normal queries that the system is expected to handle successfully. The "clean" sentences are further divided into two groups depending on the parsing results.

Table 3-1 summarizes the distribution and parsing percentage over all the spontaneous training data. The parsing coverage is 94.3% if evaluated only using the "clean" sentences. We notice that some of the sentences in the "clean & failed" category can be parsed with minor expansion of the grammar, while some are unusual and rare expressions and should not be added as rules to the grammar.

	Out-of-D	**Disfluency**	**Clean**		**Total**
			Failed	**Parsed**	
No. of Utts.	350	101	136	2266	2853
Percentage	12.3%	3.5%	4.8%	79.4%	100%

Table 3-1: Summary of parsing coverage on the orthographies of the spontaneous training utterances

In order to check the parsing quality, we examined the semantic frames from the parsing result of the "clean & parsed" category manually. Most of the semantic frames are a perfect representation of the meaning of the original sentence, and we would get the equivalent result if the corresponding English sentence was parsed. This proves that our goal of realizing a language-independent semantic representation is achieved. But there are also a few minor differences in some semantic frames. For example, some semantic frames resulting from Chinese have particles, and the singular or plural information about a noun is missing because they are not distinguished in Chinese. These differences will affect the language generation system when generating a paraphrase from the semantic frame across different languages, but they will not affect the processing function of the backend system.

3.4 Interface between Recognizer and Parser

Many times the best sentence hypothesis proposed by the recognizer is not necessary the correct answer; however, we suspect that the right one should have relatively high score and is among the top several hypotheses. So we use an N-best list as the interface between the recognizer and the parser. That is, the recognizer proposes N-best hypotheses as output; then TINA selects one of them by further applying the syntactic and semantic constraints specified in its grammar.

The following selection procedure is used for simplicity. We let TINA parse the N-best hypotheses in the order of their scores computed by the recognizer, and choose the first successfully parsed sentence. Refer to Figure 3-14 for the parsing coverage of N-best outputs for the 500 utterances in the development data set, with N varied from 1 to 10.

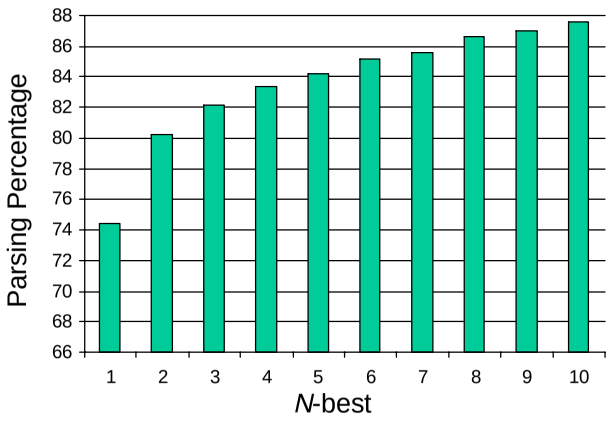


Figure 3-14: Parsing coverage on recognizer N-best hypotheses with varying N

There is a substantial improvement in terms of parsing coverage. The parsing rate on 1-best output is 74.4%, and it gets a 9.8% improvement with 5-best outputs and a 13.2% improvement with 10-best outputs. In order to check whether the improvement of parsing rate is meaningful, we examined the semantic frames from the 446 successfully parsed sentences of 10-best hypothesis, as summarized in Table 3-2. 120 of these 446 sentences are not identical to the orthography. However, 58 of these sentences result in the same semantic frames as the orthography will generate, and 46

get the basic meaning correct with mistakes in keywords. Only 16 output sentences get totally unacceptable meaning representation. This proves that the speech understanding performance is improved substantially. It also shows that sometimes we can get the correct meaning without recognizing each word in the sentence.

Parsed output	Same as orthography	Same in semantic frame	Partially same in semantic frame	Completely wrong
446	326	58	46	16

Table 3-2: Quality analysis of parsed recognizer 10-best outputs on development set data

We also examined the number of sentences without error from the recognition output as compared to the orthography. It is improved from 313 sentences with 1-best output to 348 sentences with 10-best output, a 7% improvement. We can view the grammar of the parser as a more constrained language model, so that the agrammatical errors in the recognition hypothesis can be further reduced through the selection of parsing.

3.5 Summary

In this Chapter, we briefly introduced the basic principles of the TINA parser and its application to the NL processing for the Chinese language. We described an efficient way to port grammar rules from English to Chinese. Some problems specific to the Chinese language were addressed, and solutions based on the current system were proposed. We have shown that various processing techniques provided by the TINA system can be applied to process Chinese directly or with some minor modifications. This confirms the idea that for restricted application domains, we can use the same parsing mechanisms for different languages while relying on the grammar rules to handle language specific information, provided the parsing mechanisms are complete and general enough to handle various special phenomena in different languages. The parsing performance for Chinese on the training data was reported in terms of both

coverage and quality. Experiments on an N-best list interface between the recognizer and the parser were performed. We find that applying constraints of the TINA grammar to post-process recognition results can improve performance substantially. The results also indicate that it is possible to get the sentence meaning correct even with imperfect recognition. We will give a formal evaluation on test data in Chapter 5.

Chapter 4

Natural Language Generation

Natural language generation involves generating well-formed sentences from the semantic frame representation; it is used to generate verbal *responses* to user queries as well as to provide *paraphrases* for the input utterances. The response generation capability enables the system to communicate with the user in natural language; the ability to paraphrase the input query has a variety of different usages: to help the system developers to check the appropriateness of the semantic frame resulting from the language understanding system, to reveal to the user the current status of the conversation as perceived by the system (including discourse information), and potentially to provide a translation capability.

Natural language generation is achieved by the GENESIS system [16], which uses three modules to specify linguistic rules for generation: a set of message templates, a lexicon, and a set of rewrite rules. GENESIS recursively processes pieces of a semantic frame (e.g., a topic, predicate or clause) according to these rules to form a sentence. The same generation mechanism operates for all languages, while the language-dependent rules are implemented external to the system itself; so that only a new lexicon, messages, and rewrite rules need to be developed when porting the system to a new language. We found that the process of generating correct paraphrases and responses in Mandarin Chinese was quite straightforward, and the lack of inflectional markers in the language greatly simplifies the lexicon as compared to that of English. The message templates for Chinese and English greatly differ from

each other due to the different sentence structure and word order between the two languages; however, we were able, for the most part, to utilize the GENESIS system directly to process Chinese.

In this Chapter, we first give a general description of the GENESIS system, including the specification of generation rules, and the operations of generation mechanism for response generation and paraphrasing. After that, some of the special issues in Chinese generation are addressed. Specifically, we will introduce the "topicalization" and "particle" problems in more detail, and point out some other difficulties in Chinese-English trans-lingual paraphrasing. The evaluation for the GENESIS system on unseen test data is provided in Chapter 5.

4.1 Linguistic Rules for Generation

Linguistic rules for language generation are specified in three modules: a set of message templates, a lexicon, and a set of rewrite rules. These tables incorporate the language-dependent knowledge for generation; they are also domain-specific, which means each application domain has its own set of messages, vocabulary, and rewrite rules. This feature provides a way to incorporate the domain as a high-level context indicator, which can be used to achieve context-dependent realizations of the same semantic piece for some cases, as shown in the example in Section 4.2.2. In the following, we describe each of the modules in greater detail.

4.1.1 Message Templates

The catalog of message templates is primarily used to recursively construct phrases describing topics, predicates, and clauses of a semantic frame. A message template consists of a message *name* and a sequence of one or more *word strings* and *keywords*, which describes the temporal order of the components. Both the content of the message templates and their relative ordering in the catalog play a role in deciding the temporal order of the sentence constituents, which are, for Chinese, quite different from English.

Clauses

The clauses are typically the top-level structure of the semantic frame; their role is to synthesize the phrases from the sub-level structures into a sentence-level message. Some clause templates are used for response generation; they are usually dominated by word strings, with only a few simple keywords to be filled in according to the values provided by the system manager. Table 4-1 shows some examples of clauses used for generating directions for the city guide domain. Other clause templates are used for paraphrase generation; they usually specify high-level sentence structure, with keywords augmenting the detailed substructure. Table 4-2 shows some examples of clauses used for paraphrasing for the air travel domain.

Topics

The topics typically correspond to noun phrases, and in addition to their generic type, can contain quantifiers, names, one or more predicate modifiers, and particles in the case of Chinese. Topics are synthesized by first creating a core topic noun phrase, according to either a generic topic message template or any specific topic message templates, if applicable. The predicates are then added recursively according to the "np-<*pred*>" templates corresponding to those *predicates*. The original ordering of multiple predicates within the semantic frame is actually ignored by the generation component. Instead, they are added in the order of their occurrences in the message catalog. Since Chinese is left-recursive, the predicate closer to a noun occurs earlier in the catalog. The lack of inflectional markers in Chinese obviates the difficult task of maintaining syntactic constraints of quantifiers, nouns and verbs in accordance with gender, number and case. However, the pervasive usage of particles for nouns in Chinese is a very sophisticated problem, and we address this issue in more detail in Section 4.3.2.

Predicates

The predicate category includes adjectives and prepositional phrases as well as the standard verbal predicates. In general, the surface form for a particular predicate depends upon whether or not it is contained in a clause. When modifying noun phrases, it is realized using the "np-<*pred*>" template mentioned above; when occurring in clauses, it is realized using the "<*pred*>" template directly. Some examples are shown in Table 4-2.

<*CLAUSE*>							
follow-traffic	cong2	(:NAME	na4 er5)	chu1 fa1,	yan2 zhe5	:TOPIC	zou3.
	from	(:NAME	there)	start,	along	:TOPIC	go.
cont-thru	chuan1 guo4	:TOPIC	shang4	de5	:LIGHTS .		
	go across	:TOPIC	on	+s	:LIGHTS .		
finish-left	(:NAME	ta1)	jiu4 zai4	zuo3 shou3	:TOPIC	:ADDRESS	hao4.
	(:NAME	it)	at	left-hand	:TOPIC	:ADDRESS	number.
<*TOPIC*>							
topic	:QUANTIFIER	:PARTICLE	:NOUN_PHRASE				
lights	:NUM_LIGHTS	ge4	:LIGHT				
	:NUM_LIGHTS	[particle]	:LIGHT				

Table 4-1: Example message templates for the city guide domain

<*CLAUSE*>				
display	xian3 shi4	:TOPIC .		
	show	:TOPIC .		
existential	you3 mei2 you3	:TOPIC	:PREDICATE ?	
	have not have	:TOPIC	:PREDICATE ?	
<*TOPIC*>				
topic	:QUANTIFIER	:PARTICLE	:NOUN_PHRASE	
<*PREDICATE*>				
to	:PREDICATE	:TOPIC		
np-to	:PREDICATE	:TOPIC	de5	:NOUN_PHRASE
	:PREDICATE	:TOPIC	+s	:NOUN_PHRASE
np-from	:PREDICATE	:TOPIC		:NOUN_PHRASE

Table 4-2: Example message templates for the air travel domain

4.1.2 Lexicon

The lexicon's main role is to specify the surface form of a semantic frame entry, including the construction of inflectional endings (gender, case, number, etc.). Since the Chinese language does not have these inflectional changes, the lexicon is simply a direct translation of the semantic entry to the corresponding Chinese word, as shown in Table 4-3 and Table 4-4. A lexical entry consists of a *semantic entry*, the *part of speech* for this entry, and its *surface form* in the target language. Because the semantic frame is specified using English, the lexicon looks very much like a simple English Chinese dictionary. However, since the semantic frame was originally designed for English, the map between the semantic entry and Chinese vocabulary is not always straightforward. Sometimes one semantic entry could have different Chinese translations depending on context. For example, prepositions can correspond to different Chinese words in different circumstances, or need to be dropped in some cases. We do not have a general solution for all situations; however, we were able to solve most of them case by case by writing generation rules properly. We will return to this issue later in the following sections. Sometimes, there is no obvious Chinese correspondence for a semantic entry. For example, the word "easy", as in "make an easy left turn", has to be mapped very specifically to a word that would fit in the corresponding Chinese phrase; since we are generating sentences within limited domains, this solution seems to be adequate for our current system.

massachusetts	N	ma2 sheng3
avenue	N	da4 jie1
light	N	hong2 luu4 deng1

Table 4-3: Example lexical entries for the city guide domain

flight	N	hang2 ban1
new_york	N	niu3 yue1
beijing	N	bei3 jing1
from	PREP	cong2
to	PREP	fei1 wang3

Table 4-4: Example lexical entries for the air travel domain

4.1.3 Rewrite Rules

The rewrite-rules are intended to capture surface phonotactic constraints and contractions. For example, rewriting "a other" into "another" for English. We used rewrite rules in a similar way for Chinese. For example, the phrase "there are 了 ..." corresponds to "you3 了 ..." (have 了 ...) in Chinese. However, "you3 *wu2* ..." (have *no*) does not constitute a valid phrase for Chinese (although it is quite common to say "there is no..." in English). We allow it to be generated in the same way as "you3 了 ...", and then use a simple rewrite rule to change "you3 wu2" into "mei2 you3" to form a natural phrase for Chinese. Rewrite rules are also applied in selectively deleting prepositions, as shown in the example in Section 4.3.1.

4.2 Generation Processes

The generation processes for system responses and paraphrases are slightly different. In this section, we describe each generation process in more detail, with illustrations of examples.

4.2.1 Response Generation

The response semantic frame is created by the specialized domain servers, and is usually derived from the input frame and modified to reflect the outcome of the database query. The system response generation is basically keyword-driven. The clause templates usually determine the main sentence structure, with only a few keywords to be filled in according to the values returned from the domain servers.

Refer to Figure 4-1, which shows an example of generating verbal directions in Chinese. There are three clauses involved in this example, namely *follow-traffic*, *cont-thru*, and *finish-left*, the message templates of which are shown in Table 4-1. The keywords are usually simple noun phrases. They are synthesized either using the corresponding message templates and lexical entries, such as the keyword ":TOPIC"; or simply by plugging in the values returned from the domain server, such as the

keyword ":ADDRESS"; or a combination of both, such as the keyword ":LIGHTS".

cong2	*M I T*	chu1 fa1,	yan2 zhe5	*ma2 sheng3 da4 jie1*		zou3.	
from	*M I T*	start,	along	*Massachusetts Avenue*		go.	
chuan1 guo4	*ma2 sheng3 da4 jie1*		shang4	de5	*9*	ge4	*hong2 luu4 deng1.*
go across	*Massachusetts Avenue*		on	+s	*9*	[particle]	*traffic light.*
Harvard	jiu3 zai4	zuo3 shou3	*ma2 sheng3 da4 jie1*		*1300*	hao4.	
Harvard	at	left-hand	*Massachusetts Avenue*		*1300*	number.	

Figure 4-1: An example of response generation for directions in the City Guide domain, which produces the Chinese equivalent of "Starting from MIT, follow the traffic on Massachusetts Avenue. Continue through 9 lights on Massachusetts Avenue. Harvard will be at 1300 Massachusetts Avenue on your left."

4.2.2 Paraphrase Generation

The semantic frame for the paraphrase is produced by the language understanding system TINA, which is the meaning representation of the input query. The paraphrase semantic frame usually has a more linguistically motivated hierarchical structure; and the proper temporal order of the various constituents in the reconstructed sentence relies more on the structures of various message templates and their relative order in the message catalog.

```
{c display
    :topic {q flight
                :pred {p from
                            :topic {q city
                                        :name "new york" } }
                :pred {p to
                            :topic {q city
                                        :name "beijing" } } }
    :domain "AirTravel" }
```

Figure 4-2: Semantic frame for the Chinese sentence "Show from New York fly to Beijing +s flight."

We use the example shown in Figure 4-2 to illustrate the generation process. The *domain* entry in the semantic frame specifies "air travel", so the messages and lexicon

shown in Table 4-2 and Table 4-4 are used for reconstruction. The top-level *clause* type is "display", which corresponds to the sentence template "show :TOPIC ." The *topic* entry in the semantic frame indicates "flight", with two *predicates*, "from" and "to". The order of these two predicates in the semantic frame is ignored, and the actual temporal order is determined by their order in the message templates. In constructing the topic, the core topic noun phrase "flight" is first synthesized; then the predicate "to" is added to the structure (because *np-to* precedes *np-from* in the messages) by looking up the template "np-to", yielding the phrase "to beijing +s flight". After that, the predicate "from" is added to the structure (the noun-phrase is "to beijing + flight" instead of "flight" at this stage), generating the phrase "from new york to beijing +s flight" which is the final ":TOPIC." By plugging the :TOPIC into the sentence, we get "show from new york to beijing +s flight," which is exactly the same as the input sentence. The word-by-word English translations of message templates were used here for illustration purpose. Of course, the message templates with Chinese word strings were used in the actual process, and all the semantic entries were translated into Chinese words according to the lexicon; thus yielding a natural Chinese sentence "xian3 shi4 cong2 niu3 yue1 fei1 wang3 bei3 jing1 de5 hang2 ban1."

Usually the prepositions in Chinese have different surface forms depending on the context. For example, in the English phrases "directions from MIT to Harvard" and "flights from Boston to Beijing", the prepositions "from" and "to" are the same in both cases. However, in the corresponding Chinese phrases, the "to" in the first case is a generic form "dao4" (to), while the "to" in the second phrase implies "fei1 wang3" (fly to). It is very difficult to achieve a general solution for this phenomenon. However, we utilized the domain-specific feature to reach a very simple solution for our application domains. Notice that the predicate "to" in the air travel domain implies flight destination, so we simply specify the surface form for "to" as "fei1 wang3" (fly to) in the lexicon for the air travel domain; while the "to" in the city guide domain is realized differently as "dao4"(to).

4.3 Issues in Chinese Generation

Movement phenomena are one of the most difficult aspects of paraphrase generation. Although the wh-questions in Chinese have the normal word order, the "are there ..." type of questions in Chinese usually involve long-distance movement; and it would be highly desirable if we can recover the same word order as the input sentence when generating a paraphrase from the semantic frame. Another challenge is brought by the use of particles to accompany nouns in Chinese. It is very difficult to decide which particle to use under what circumstances due to the complex context effects involved; yet the particles have to be generated properly in order to obtain a natural Chinese sentence. In this section, we address these issues in detail; some difficulties in trans-lingual paraphrasing (translation) will also be described briefly.

4.3.1 Topic Movement

In Chapter 3, the long-distance movement or "topicalization" phenomenon was addressed for language understanding. It was mentioned that TINA was able to tag a special mark "FOCUS" to the moved topic to facilitate correct paraphrasing. In the following, we continue to discuss the "topicalization" problem for paraphrase generation in more detail. We use the same example as in Section 3.2.1 to illustrate the generation process. The semantic frame is repeated here in Figure 4-3 for convenience.

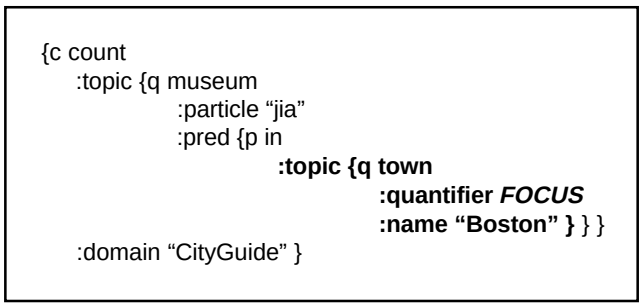


Figure 4-3: Semantic frame for an example Chinese sentence with long-distance movement

The message templates, lexical entries, and rewrite rules needed in this example

are shown in Table 4-5. The ":TRACE" in the *count* message template is critical to achieving topic movement. The *trace* mechanism in the GENESIS system works as follows: if the *part of speech* of a semantic entry is assigned as "TRACE" in the lexicon, then the semantic piece it directly belongs to is substituted into the ":TRACE" keyword in the clause template, regardless of its original position in the semantic frame. In our example, the "FOCUS" is assigned the "TRACE" property in the lexicon, so the *topic* "town" becomes ":TRACE", thus disregarding its original relationship with the *predicate* "in". It is synthesized as a simple topic, yielding the noun phrase "Boston" as ":TRACE" at the sentence initial position. Since the clause contains the "PREP", the *predicate* "in" is synthesized as "zai4" according to the *in* message template; and the *topic* "museum" is simply synthesized as the core noun phrase without any predicate.

The sentence is still not perfect so far, because the preposition "in" needs to be dropped during "topicalization". We achieve this by the collaboration of a *deletion mark* "*del*" and rewrite rules, as shown in Table 4-5. The preposition "zai4" for this case always appears together with "*del*", and they are both deleted because of the rewrite rule (" *del* zai4 " $\Rightarrow$ " "). This rule will not affect "zai4" in other cases where it is not tagged with "*del*".

Perhaps a more obvious solution would be to simply delete the ":PREP" keyword from the clause template. However, there are cases where prepositions need to be kept in "topicalization". For example, "near" is usually kept intact, and has the surface form "fu4 jin4" (vicinity). Since there is no rewrite rule to delete "*del* fu4 jin4", only "*del*" will be deleted from the final sentence, leaving the desired preposition "fu4 jin4" intact. This method can actually serve as a general "selective deletion" mechanism and is applicable in many other situations.

4.3.2 Particles

One aspect of the Chinese language that is quite different from English is the use of particles to accompany nouns. The concept of particle is analogous to "a *flock* of sheep" in English, except that they are far more pervasive in the language. Thus "a

count	:TRACE	*del*	:PREP	you3	duo1 shao3	:TOPIC ?
	:TRACE	*del*	:PREP	have	how many	:TOPIC ?
topic	:QUANTIFIER :PARTICLE :NOUN_PHRASE					
in	:PREDICATE :TOPIC					

focus	TRACE	" "
museum	N	bo2 wu4 guan3
boston	N	bo1 shi4 dun4
in	PREP	zai4

" *del* zai4 "	" "
" *del* "	" "

Table 4-5: Message templates, lexical entries, and rewrite rules used in a long-distance movement example for language generation

bank" becomes "a <particle> bank", "this bank" becomes "this <particle> bank, and "these banks" could become "these several <particle> bank," etc. The exact realization of the particle depends on the class of the noun, and there is a fairly large number of possibilities. The situation is further complicated by the fact that the particle does not necessarily always accompany the noun, and the exact usage is very hard to summarize due to the complex context effects.

Our current approach to solving this problem is to preserve the particle information in the semantic frame. For example, in the semantic frame shown in Figure 4-3, the *topic* "museum" has a *particle* "jia1" which is obtained from parsing the input utterance. Using the message template *topic* in Table 4-5, the particle is then recovered in the paraphrase. Thus the issue of when to use which particle is avoided by simply keeping consistency with the input sentence. However, this method would not be able to produce trans-lingual paraphrases that are well-formed. For example, no particle information is available in a semantic frame from an English sentence, so the resulting Chinese paraphrase will not be able to recover particles properly.

4.3.3 Difficulties for Trans-lingual Paraphrasing

Our approaches to solving topic movement and particle generation both rely on incorporating language-dependent syntactic information in the semantic frame. Although they are inconsequential to the "common meaning representation" approach for our multilingual system, the semantic frame representation is not totally language-independent concerning syntactic information.

There are also some other inconsistencies between the semantic frames of Chinese and western languages. For example, the Chinese language lacks inflectional changes; so the gender, number and case information is usually missing in the semantic frame. It is not a problem in Chinese generation; however, it could be problematic in generating English sentences, which usually relies on the number information of a noun provided in the semantic frame for proper paraphrasing.

The approach of relying on the semantic frame to incorporate some language-dependent information for generation provides an easy solution to many difficult generation problems, and it seems to be adequate for the application in GALAXY, because the system usually operates in one common input/output language, although the language can vary. However, more general solutions need to be implemented in order to achieve *trans-lingual interaction*.

4.4 Summary

In this chapter, we described the basic principles of the GENESIS system and its application in Chinese language generation. GENESIS uses the same generation mechanism for all languages, with the language-dependent generation rules implemented externally as a lexicon, a set of message templates, and a set of rewrite rules. Overall, we found that the process of generating correct responses and paraphrases in Mandarin Chinese was quite straightforward, and we were able to utilize the GENESIS framework directly for the most part. However, some special language phenomena are difficult to handle only by the generation modules, and we have to rely on the semantic frame to incorporate some language-dependent syntactic information. This seems to be an

adequate and easy solution for the application in GALAXY, which usually operates with consistent input/output languages. However, more general alternatives need to be investigated to achieve proper translation.

Chapter 5

Evaluation

The performance of an interactive spoken language system could depend on all the individual component in the system and their interactions with each other, as well as human factors. Designing a good evaluation method to assess the system objectively is by itself a very complicated research issue [18].

However, since this thesis involves providing a Mandarin speech interface to the existing conversational system GALAXY, we can simplify the evaluation task by only evaluating the relevant parts of the system. Specifically, we are more interested in the performance related to the speech recognition, language understanding, and response generation components, and the evaluation metrics could be chosen such that the influences from the system manager, discourse component, and database access are excluded. In this chapter, we describe our evaluation methodology and report the results.

5.1 Evaluation Methodology

5.1.1 Data

Spontaneous utterances from 3 female and 3 male speakers were set aside as test data. They are all native speakers of Mandarin Chinese. Half of them are from the northern part of China and the other half from the south. No data from these speakers have

been used in any training. Refer to Table 2-1 for a summary of the test data.

5.1.2 Methodology

Figure 5-1 shows a simplified model of the system architecture, emphasizing the major signal flow. The system manager, discourse component, and database are treated as a black box referred to as "backend" in the figure. Since we are not interested in the performance of the system backend, the evaluation is divided into two parts. On the input side, the performance from speech signal to semantic representation is examined. On the output side, the performance from semantic frame to natural sentence is examined. Four experiments are designed to evaluate the input and output stages, as well as each individual component.

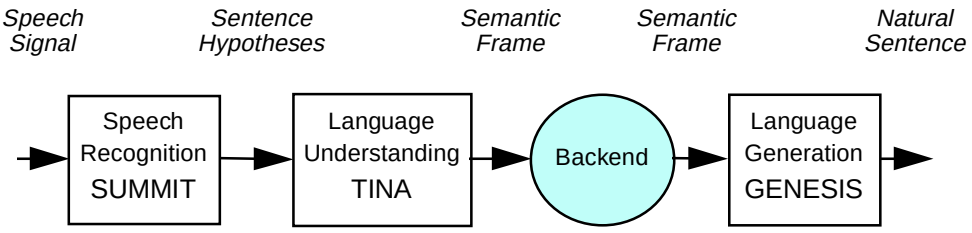


Figure 5-1: System architecture model for evaluation

Speech Recognition

The speech recognition performance can be naturally separated from the influences of the other parts of the system. It is evaluated using the speech waveform in the test data set, and the standard *word error rate* and *sentence error rate* measurements of recognizer 1-best hypothesis are used as evaluation metrics. The performance here is based on the best hypothesis only; in practice, an N-best list is filtered by the parser, so the recognition performance could be better.

Language Understanding

Language understanding concerns obtaining the meaning representation from the input sentence (text form). In order to evaluate the stand-alone performance of the TINA system, we use the orthography of the test set as evaluation data to exclude the influence from the recognizer, and the input sentences are treated as stand-alone queries to exclude discourse influences. The *parsing coverage* and *quality of the semantic frame* are used as evaluation metrics.

Speech Understanding

Speech understanding combines speech recognition and language understanding. It corresponds to the input stage from speech signal to semantic frame representation. Since it is possible for the parser to derive a correct or partially correct meaning from the input utterance even with some errors in recognition, this evaluation can not be substituted by the simple combination of the previous two evaluations. We described an N-best list interface between the recognizer and the parser in Chapter 3. A 10-best list is used in the final system configuration, and we use the same *parsing coverage* and *quality of semantic frame* as evaluation metrics.

Language Generation

Language generation involves forming a natural sentence from a semantic frame representation. There are two types of generation in the system: form a paraphrase for the input sentence, or generate the system response in natural sentences. For the first task, we use the correct semantic frames resulting from TINA as the evaluation data, and the "naturalness" of the resulting sentence is used as the evaluation metric. The second generation task is relatively easy, because the system responses are a very limited set and the generation is through a key-word driven approach. We feel it is not necessary to provide separate evaluation based on the observation that the system responses behaved well during data collection.

5.2 Results and Analysis

5.2.1 Speech recognition

Table 5-1 summarizes the recognition performance on both the test set and development set. There is only slight degradation from development set performance to unseen test set performance.

	No. of Utts.	**WER**	**SER**
Dev. Set	500	9.1%	37.4%
Test Set	274	10.8%	39.1%

Table 5-1: Summary of recognition performance on test set data

5.2.2 Language Understanding

Table 5-2 shows the language understanding performance evaluated using the orthography of the 274 utterances in the test data set. A sentence can either be parsed by TINA or fail to parse. For the parsed sentences, the resulting semantic frames are further classified into three disjoint categories depending on the quality of the semantic frame. The *perfect* category means that the semantic frame represents the meaning of the input sentence correctly; the *acceptable* category means that the semantic frame gets the basic meaning of the input sentence correctly, but with errors in some keywords; and the *wrong* category means that the semantic frame gets the basic meaning of the sentence wrong. About 16% of the sentences failed to parsed, and among the parsed sentences, most of the semantic frames belong to the *perfect* category.

In order to examine the performance of the parser more carefully, we analyzed the 44 sentences that failed to parse. We found that 30 queries in this category were "out-of-domain" or "disfluent" (as defined in Chapter 3), while the remaining 14 sentences were mostly reasonable queries. Figure 5-2 shows some of the failed sentences.

	Parsed			
	Perfect	**Acceptable**	**Wrong**	**Failed**
No. of Utts.	220	8	2	44
Percentage	80.29%	2.92%	0.73%	16.06%

Table 5-2: Summary of language understanding performance on test set data

Out-of-domain:	m i t	li2	charles he2	you3 duo1 yuan3			
	m i t	from	charles river	how far			
Out-of-domain:	bo1 shi4 dun4		zui4 hao3 de5	luu3 guan3	shi4	na3 yi1 jia1	
	Boston		best	hotel	is	which one	
Disfluent:	wu3 yue4 hua1		dao4	ha1 fo2	you3	duo1 shao3	duo1 yuan3
	Royal East		to	Harvard	have	how many	how far
Normal:	qi2 zhong1	you3	duo1 shao3	zhong1 can1 guan3			
	among these	have	how many	Chinese restaurant			

Figure 5-2: Examples of test set sentences that failed to parse

5.2.3 Speech Understanding

A 10-best list is used as the interface between SUMMIT and TINA in the final system configuration. Table 5-3 shows the speech understanding performance. The results on the 1-best recognizer hypothesis are also included for comparison.

		Parsed			
		Perfect	**Acceptable**	**Wrong**	**Failed**
1-best	No. of Utts.	171	19	7	77
	Percentage	62.41%	6.93%	2.56%	28.10%
10-best	No. of Utts.	191	23	15	45
	Percentage	69.71%	8.39%	5.48%	16.42%

Table 5-3: Summary of speech understanding performance on test set data

The percentage of sentences yielding "perfect" semantic frames is significantly increased with the 10-best list, which means that the post-processing of the parser on

recognition hypotheses is effective in selecting more plausible candidates using grammatical constraints. However, there are also significant increases in the "acceptable" and "wrong" categories. This is because the grammar is not always constrained to incorporate only the "right" rules; sometimes the rules are relaxed to parse ill-formed sentences as well to improve parsing coverage, and many nonsense sentence hypotheses proposed by the recognizer might seem plausible to the over-generalized rules. The trade-off between parsing coverage and constraints needs to be carefully balanced in writing a grammar, and this is a very challenging task to accomplish.

Comparing Table 5-3 and Table 5-2, the gap of understanding performance between speech input (with a 10-best recognizer output) and text input is only about 10%, which means that the degradation due to speech recognition errors is relatively small.

5.2.4 Language Generation

The evaluation for the language generation performance is somewhat subjective. The quality of the resulting sentence is judged to be natural or unnatural by a human evaluator. Since the system response generation is a well covered set, we did not perform evaluation for that part. However, it is verified that all the sentences generated from the 220 "perfect" semantic frames referred in Table 5-2 can be considered as a "natural" paraphrases of the corresponding orthography.

Chapter 6

Summary and Future Work

6.1 Summary

This thesis concerns porting the GALAXY conversational system to Mandarin Chinese. Speech recognition, language understanding, and language generation components were developed for Mandarin; large amounts of Mandarin speech data were collected from native speakers for system development. Comparisons between the Chinese and English language were made in the context of system implementation. Some issues that came up in porting the recognizer, the understanding component, and the generation component to Mandarin Chinese were addressed, to make the system core more language independent.

Data collection is a significant and time consuming effort, which is absolutely necessary for the development of a high-performance system. We were able to collect more than 7,000 spontaneous and read utterances from native speakers of Mandarin Chinese for system development. The 3,100 spontaneous utterances were collected in *wizard mode*, which were used for training both acoustic and language models for recognition, and deriving and training a grammar for language understanding; the 4,200 read utterances were collected using our Web data collection facility, they were very valuable for acoustic training due to the phone-line diversity.

The probabilistic, feature-based SUMMIT recognizer was configured appropriately to carry out Mandarin speech recognition. Acoustic models, lexical items (with pro-

nunciations), and language models for the recognizer were derived for Mandarin Chinese from the training data. Since the SUMMIT system currently does not have proper mechanisms to incorporate tone recognition, we adopted the approach of performing only base syllable recognition, realizing that this leads to a greater number of potential homophones; disambiguation of homophones was performed at the level of parsing. The current vocabulary has about 1,000 words, containing both Chinese and English because of the application scenarios; about one quarter of the vocabulary are English, and each Chinese word has on average 2.1 characters.

The mixed Chinese and English vocabulary makes acoustic modeling a more challenging problem. After some experiments with various sets of phonetic units, we finally settled on the simple choice of using Chinese syllable initials and finals as phonetic units, augmented with only a few English-specific phonemes; the remaining English pronunciations were approximated by near-neighbor Mandarin equivalent. Mixtures of diagonal Gaussians were used as acoustic models; the models were seeded from English models, due to the lack of pre-existing local acoustic models for Chinese, and trained on the collected data using an iterative training process. Dialectal variations of the phonetic units are handled through the Gaussian mixtures. Several bigram models have been explored, and a class bigram is used in the current system. The recognizer achieved 10.8% word error rate and 39.1% sentence error rate on the unseen test data.

Language understanding for Chinese is processed by the TINA system, which extracts the meaning of an input sentence in the form of a semantic frame, with language-dependent knowledge specified by hand-written context-free grammar rules. The grammar rules of TINA combine semantic categories and syntactic categories; we believe that this approach is very suitable for parsing Chinese, because the lack of inflectional markers in the language makes purely syntactic approaches hard to succeed. We determined the appropriate grammar rules for each new Chinese 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. The Chinese grammar tends

to be left recursive, which is incompatible with TINA's top-down control flow; we have to be careful to avoid writing such rules and use non-recursive equivalents instead. Aside from that, the parser appears to be competent in processing Chinese. Tokenization and homophone problems are handled implicitly by the grammar rules, and many mechanisms can be applied directly for Chinese or with minor changes, for example, the trace mechanism originally designed to process English wh-questions is successfully applied in dealing with long-distance movement associated with "topicalization" in Chinese. In formal evaluation, the parser produced correct semantic frames for about 80% of the orthographies of the test data.

Natural language generation for Chinese involves generating well-formed Chinese sentences from the semantic frame representation; it is performed by the GENESIS system. GENESIS uses three modules to specify linguistic rules for generation: a set of message templates, a lexicon, and a set of rewrite rules. The process of developing these rules for Chinese was very straightforward; and for the most part, we were able to utilize the GENESIS system directly for Chinese. There are some special language phenomena in Chinese that are difficult to handle within the current GENESIS framework, for example, the long-distance movement associated with "topicalization" and the "particles" for Chinese nouns; however, we were able to solve them by incorporating some language-dependent syntactic information in the semantic frame during parsing to assist generation. This seems to be an adequate and easy solution for the application in GALAXY, which usually operates with consistent input/output languages. GENESIS is applied to generate verbal responses to user queries as well as to provide paraphrases for the input queries. We were able to obtain good performance for both tasks; it was verified that for all the correct semantic frames resulting from parsing the test data, the paraphrase generated by GENESIS could be considered as natural.

In the final system, the recognizer interfaces with the parser through a 10-best list. Overall, the system produced reasonable responses nearly 70% of the time for the test data collected in the wizard mode, comparable in performance to its English counterpart. This demonstrates the feasibility of the design of GALAXY aimed at

accommodating multiple languages in a common framework.

6.2 Future Work

Chinese is tonal syllabic, in which each character represents a syllable with a particular tone, and words are formed by one or multiple characters, as illustrated in Figure 6-1. Considering that there are only about 416 base syllables in the entire language, which form the vocabulary of more than 6,000 words, developing an adequate syllable recognizer is feasible and highly desirable for Chinese. A recognizer based on syllables can potentially solve the unlimited-vocabulary recognition problem, it is also much easier to incorporate tone recognition into this framework because of the explicit handles for syllables. However, language modeling based on syllables is a very challenging problem to solve. We suspect that the widely used *N*-gram based language models can not provide adequate constraints for syllable transitions, and other alternatives such as hierarchical language models must be investigated to achieve high quality syllable recognition.

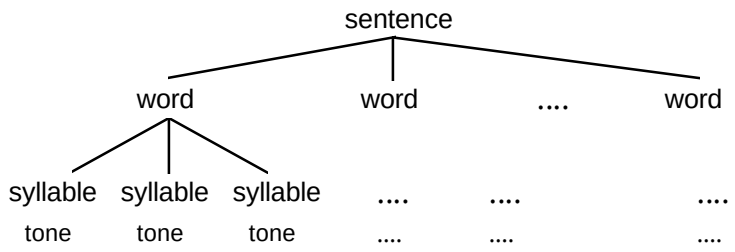


Figure 6-1: Parse tree illustrating the syllabic feature of Chinese

While our system can converse with the user in Mandarin Chinese, it often displays the information in English, due to the English-based information sources. Thus, if the user asks for the weather for Beijing China, it says (displays) "Here is the weather for Beijing" in Chinese, and then shows the weather forecast in English. We have actually begun the process of translating on-line weather reports from English to Chinese, so that the *information itself*, and not just the interface for accessing the information, will be provided to the user in the preferred language.

We also need to obtain a Mandarin speech synthesizer to incorporate into the system to make it truly *conversational*.

Bibliography

[1] J. Allen. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc., 1994.

[2] B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, and G. Lazzari. Multilingual person to person communication at IRST. In *Proc. ICASSP*, pages 91–94, 1997.

[3] A. Bayya, M. Durian, L. Meiskey, R. Root, R. Sparks, and M. Terry. VOICEMAP: A dialogue-based spoken language information access system. In *Proc. ICSLP*, 1994.

[4] M. Blomberg, R. Carlson, K. Elenius, B. Granstrom, J. Gustafson, S. Hunnicutt, R. Lindell, and L. Neovius. An experimental dialogue system: WAXHOLM. In *Proc. Eurospeech*, pages 1867–1870, 1993.

[5] T. Bub and J. Schwinn. VERBMOBIL: The evolution of a complex large speech-to-speech translation system. In *Proc. ICSLP*, pages 2371–2374, 1996.

[6] J. Chang. Speech recognition robustness to microphone variations. Master's thesis, Massachusetts Institute of Technology, 1995.

[7] K. J. Chen. A model for robust chinese parser. *International Journal of Computational Linguistics & Chinese Language Processing*, 1(1), 1996.

[8] D. Clementino and L. Fissore. A man-machine dialogue system for speech access to train timetable information. In *Proc. Eurospeech*, pages 1863–1866, 1993.

[9] W. Eckert, T. Kuhn, H. Niemann, and S. Rieck. A spoken dialogue system for German intercity train timetable inquiries. In *Proc. Eurospeech*, pages 1871–1874, 1993.

[10] G. Flammia, J. Glass, M. Phillips, J. Polifroni, S. Seneff, and V. Zue. Porting the bilingual VOYAGER system to Italian. In *Proc. ICSLP*, 1994.

[11] Y. Gao, H. Hon, Z. Lin, G. Loudon, S. Yogananthan, and B. Yuan. TANGERINE: A large vocabulary Mandarin dictation system. In *Proc. ICASSP*, pages 77–80, 1995.

[12] J. Glass. *Finding acoustic regularities in speech: applications to phonetic recognition*. PhD thesis, Massachusetts Institute of Technology, 1988.

[13] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *Proc. ICSLP*, 1996.

[14] J. Glass, G. Flammia, D. Goodline, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue. Multilingual spoken language understanding in the MIT VOYAGER system. *Speech Communication*, 17:1–18, 1995.

[15] J. Glass, D. Goodine, M. Phillips, S. Sakai, S. Seneff, and V. Zue. A bilingual VOYAGER system. In *Proc. Eurospeech*, pages 2063–2066, 1993.

[16] J. Glass, J. Polifroni, and S. Seneff. Multilingual language generation across multiple domains. In *Proc. ICSLP*, 1994.

[17] 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. In *Proc. ICSLP*, 1994.

[18] D. Goodian, L. Hirschman, J. Polifroni, S. Seneff, and V. Zue. Evaluating interactive spoken language systems. In *Proc. ICSLP*, 1992.

[19] H. Hon, B. Yuan, Y. Chow, S. Narayan, and K. Lee. Towards large vocabulary Mandarin Chinese speech recognition. In *Proc. ICASSP*, pages 545–548, 1994.

[20] E. Hurley, J. Polifroni, and J. Glass. Telephone data collection using the World Wide Web. In *Proc. ICSLP*, 1996.

[21] W. Hutchins and H. Somers. *An Introduction to Machine Translation*. Academic Press, 1992.

[22] A. Lavie, A. Waibel, L. Levin, D. Gates, M. Gavalda, T. Zeppenfeld, P. Zhan, and O. Glickman. Translation of conversational speech with JANUS-II. In *Proc. ICSLP*, pages 2375–2378, 1996.

[23] R. Lyu, L. Chien, S. Hwang, H. Hsieh, R. Yang, B. Bai, J. Weng, Y. Yang, S. Lin, K. Chen, C. Tseng, and L. Lee. Golden Mandarin(III): A user-adaptive prosodic-segment-based Mandarin dictation machine for Chinese language with very large vocabulary. In *Proc. ICASSP*, pages 57–60, 1995.

[24] H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao, J. Polifroni, S. Seneff, and V. Zue. WHEELS: A conversational system in the automobile classifieds domain. In *Proc. ICSLP*, pages 542–545, 1996.

[25] T. Morimoto, T. Takezawa, Yato F., S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu. ATR's speech translation system: ASURA. In *Proc. Eurospeech*, pages 1295–1299, 1993.

[26] M. Oerder and H. Aust. A realtime prototype of an automatic inquiry system. In *Proc. ICSLP*, pages 703–706, 1994.

[27] D. Pallett, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Pryzbocki. 1993 benchmark tests for the ARPA spoken language program. In *Proc. DARPA Speech and Natural Language Workshop*, pages 49–74, 1994.

[28] J. Peckham. Speech understanding and dialogue over the telephone: an overview of the ESPRIT SUNDTAL project. In *Proc. DARPA Speech and Natural Language Workshop*, pages 14–27, 1991.

[29] S. Seneff. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1):61–86, 1992.

[30] H. Wang, J. Shen, Y. Yang, C. Tseng, and L. Lee. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data. In *Proc. ICASSP*, pages 61–64, 1995.

[31] J. Yang and Y. Lee. Toward translating Korean speech into other languages. In *Proc. ICSLP*, pages 2368–2370, 1996.

[32] V. Zue. Human computer interactions using language based technology. In *Proc. the International Symposium on Speech, Image Processing & Neural Networks*, pages 123–125. IEEE, 1994.

[33] V. Zue, J. Glass, D. Goddeau, D. Goodine, C. Pao, M. Phillips, J. Polifroni, and S. Seneff. PEGASUS: A spoken dialogue interface for on-line air travel planning. *Speech Communication*, 15:331–340, 1994.

[34] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. The VOYAGER speech understanding system: A progress report. In *Proc. DARPA Speech and Natural Language Workshop*, pages 51–59, 1989.

[35] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. The VOYAGER speech understanding system: Preliminary development and evaluation. In *Proc. ICASSP*, pages 73–76, 1990.

[36] V. Zue, J. Glass, D. Goodine, M. Phillips, and S. Seneff. The SUMMIT speech recognition system: Phonological modeling and lexical access. In *Proc. ICASSP*, pages 49–52, 1990.

[37] V. Zue, J. Glass, M. Phillips, and S. Seneff. Acoustic segmentation and phonetic classification in the SUMMIT system. In *Proc. ICASSP*, pages 389–392, 1989.

[38] V. Zue, S. Seneff, J. Polifroni, H. Meng, and J. Glass. Multilingual human-computer interactions: From information access to language learning. In *Proc. ICSLP*, pages 2207–2210, 1996.