High-quality Speech Translation for Language Learning

Chao Wang and Stephanie Seneff
MIT Computer Science and Artificial Intelligence Laboratory
Stata Center, 32 Vassar Street
Cambridge, MA 02139, USA
{wangc,seneff}@csail.mit.edu

Abstract

In this paper, we describe a translation framework aimed at achieving high-quality speech translation within restricted conversational domains. Towards this goal, we developed an interlingua-based approach, in which a generation-based method is augmented with an example-based method to improve system robustness, even with imperfect inputs due to speech recognition errors. The framework is integrated into a dialogue-based language tutoring system, to provide immediate translation assistance to students during the dialogue interaction. We evaluated the translation quality within a weather information domain configured for native English speakers practicing Mandarin Chinese. We achieved perfect or acceptable translations for 94.3% of the manual transcriptions of a test set of 695 spoken queries, and 90.2% on automatic speech recognition outputs.

1 Introduction

The main components to language learning are reading, writing, listening and speaking. While it is possible for diligent students to gain adequate proficiency in the first three areas, the goal of improving conversational skills cannot be achieved by simply working hard. This is mainly due to the lack of a proper environment and adequate opportunity to practice speaking. Dialogue systems can potentially change the situation by providing an entertaining and non-threatening conversational environment (Seneff et al., 2004). A critical technology in our vision is the ability to generate high-quality translations for speech inputs in the native language, to provide students with immediate assistance when they have difficulty expressing themselves in the new language. With this assistance, their conversation with the computer can carry on even before they have sufficient proficiency in the foreign language.

Speech-to-speech translation is a challenging problem, due to poor sentence planning typically associated with spontaneous speech, as well as errors caused by automatic speech recognition. Most speech translation systems reported in the literature operate within more or less restricted domains (Levin et al., 2000; Frederking et al., 2002; Gao et al., 2002; Rayner and Bouillon, 2002). Many are based on the interlingua approach to translation; however, systems differ in their linguistic complexity. Knowledge-lean statistical machine translation approaches are nearly universally embraced for the task of unrestricted text translation (Koehn et al., 2003), perhaps because it is more difficult to effectively exploit knowledge in the broad domain. In restricted domains, rule-based and statistical-based approaches clearly show different strengths and weaknesses, which makes them complement each other nicely.

Our translation framework adopts the interlingua approach and is integrated with our dialogue system development via a shared meaning representation which we call a *semantic frame*. Given an input sentence, a parse tree is derived and critical syntactic relations and semantic elements in the parse tree are extracted. The resulting semantic frame can be used to generate key-value (KV) information for the dialogue system, and to generate a sentence in the original language (paraphrasing) or in a different language (translation). The generation is controlled by a set of rules and a context-sensitive lexicon, which can be fine-tuned to achieve high quality. We adopt a knowledge-rich approach in both the parsing and generation components, while emphasizing portability of the grammar and generation rules to new domains (Rayner and Carter, 1997).

Our dialogue tutoring system employs two grammars, one to parse the native language (L1) for translation, and one to parse the foreign language (L2) for dialogue processing. We can make use of the L2 grammar to achieve some "quality assurance" on the translation outputs. If the generated translation failed parsing under the L2 grammar, we resort to an example-based method in which semantic information encoded as key-value pairs is used to look up a pre-compiled L2 corpus for a suitable candidate. If both methods failed, the system will prompt the student to rephrase. We think that a null output is perhaps better than an erroneous one, given the intended use of our system. The example-based mechanism complements the rule-based generation in that it tends to be more robust for ill-formed inputs (Levin et al., 2000; Frederking et al., 2002).

In the remainder of the paper, we first describe our interlingua representation, which is derived by parsing the input sentence. Sections 3 and 4 describe the generation-based method and the example-based method respectively. Empirical results on manually

and automatically derived speech transcriptions are reported in Section 5.

2 Parsing and the Interlingua

The first step in our translation procedure is to derive an interlingua representation of the input sentence, which is a structured object that hierarchically encodes the relationships among major syntactic constituents of the sentence. We use the TINA system (Seneff, 1992), which utilizes a context-free grammar to define allowable utterance patterns within each domain, augmented with a feature unification mechanism to enforce agreements and handle movement phenomena. The system supports a probability model which is acquired by tabulating frequency counts on a large corpus of parsed data.

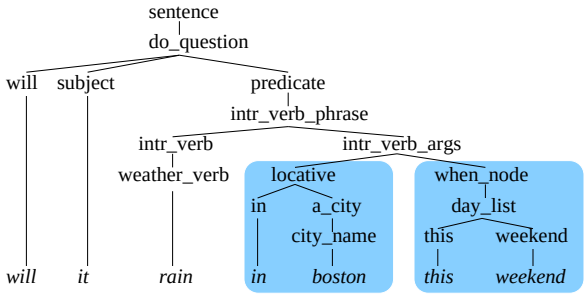
Constructing a well-formed grammar for each conversational domain (lesson plan) is a time-consuming process that requires either large amounts of labelled in-domain data to automatically induce and train a grammar (Collins, 1997), or linguistic expertise to compensate for the lack of data. A practical compromise is to induce a shallow grammar from coarsely (and possibly automatically) labelled data (Tang et al., 2002), which usually results in a relatively flat semantic structure. Such a representation, while acceptable for dialogue processing, is generally not adequate for deriving an accurate translation of the input.

Our solution is to construct a generic core grammar that is linguistically rich and easily adaptable for different domains within the dialogue interaction scenario. The grammar is induced semi-automatically, making heavy use of existing resources such as several grammars developed for various applications (Zue et al., 2000; Seneff and Polifroni, 2000), as well as speech corpora collected using those systems. The core grammar contains top-level rules specifying detailed syntactic structure, with part-of-speech nodes to be expanded into words or word classes specific to an domain. Rules for general semantic concepts such as dates and times are organized into sub-grammars that are easily embedded into any domains. Phrases for sub-grammars are also extracted from existing speech transcripts, which can be re-used to train a statistical language model for the new domain (Seneff et al., 2003).

Figure 1 illustrates a parse tree and the corresponding semantic frame derived using our core English grammar adapted for the weather domain. The sub-grammars to parse the location phrase "in Boston" and the date phrase "this weekend" (highlighted in rectangular shades in the figure) are directly provided by the core grammar. Only domain-specific nouns and verbs (e.g. "rain" as a "weather_verb") need to be entered by a developer.

3 Natural Language Generation

To generate well-formed strings in L2, we utilize the GENESIS language generation framework (Baptist and Seneff, 2000). It works from a lexicon which provides context-dependent word-sense surface strings for each vocabulary item, along with a set of recursive rules that specify the ordering of constituents in the generated string. Variability in the surface form can be achieved by randomly selecting among alternative rules and lexical entries. For example, the English example in Figure 1 can be realized as the Chinese "statement + question-particle" construct or the "A-not-A" construct, with additional permutation on the ordering of the date and location phrases. This is useful not only to the language student, but also to the system, since we can generate a rich set of Chinese sentences for training the language models of the speech recognizer.




```
{c verify
   :auxil "will"
   :topic {q pronoun
            :name "it" }
   :pred {p rain
            :pred {p locative
                     :prep "in"
                     :topic {q city
                              :name "boston" }}
            :pred {p temporal
                     :topic {q weekday
                              :quantifier "this"
                              :name "weekend" }}}}
```

Figure 1: Parse tree and semantic frame of an example sentence "*Will it rain in Boston this weekend?*"

GENESIS has recently been enhanced to include a preprocessor stage (Cowan, 2004), which handles the transfer step in the translation process. It augments the frame with syntactic and semantic features specific to the target language, for example, deciding between definite and indefinite articles for noun phrases translated from Mandarin to English. In rare cases, the *structure* of the frame must be transformed, to handle situations where a concept is expressed very differently in the two languages, for instance "what is your name?" translating literally to "You called what name?" in Mandarin.

The generation rules can be fine-tuned by experts to produce high-quality outputs on a set of development data. However, when the input deviates from the expected patterns, either due to novel linguistic constructs or caused by speech recognition errors, the rule-based generation module could produce ill-formed outputs. In order to prevent erroneous translations from confusing the student (who is trying to learn the language), we use the L2 gram-

mar developed for the dialogue system to ensure that the translation output is a legitimate sentence. We noticed that the grammar would occasionally reject good translations, due to coverage gaps in the L2 grammar. However, we think that this is a desirable feature, because those sentences will nevertheless fail in subsequent system processing, if the students choose to imitate them during their conversation with the system.

Figure 2 summarizes the generation-based translation procedure, configured for the scenario of a native English speaker learning Chinese. As demonstrated in the figure, this procedure can be used to automatically produce a key-value indexed Chinese corpus from a collection of English sentences, to serve as a translation memory for the example-based method. The translation table can also be augmented with any available original Chinese data, in which case the KV index can be derived using the Chinese grammar for parsing. Details of the KV representation are discussed in the next section.

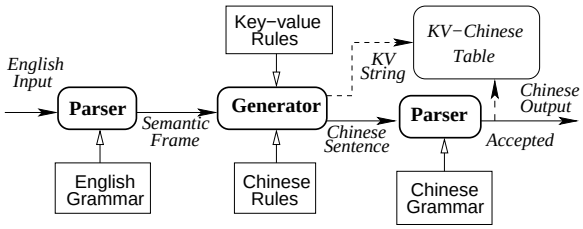


Figure 2: Schematic diagram for generation-based translation method. Note: KV = Key Value.

4 Example-based Translation

The example-based approach requires a collection of pre-existing translation pairs and a retrieval mechanism to search the translation memory. Similarity can be based on parse trees (Sato, 1992), complete sentences (Veale and Way, 1997), or words and phrases (Brown, 1999; Levin et al., 2000; Marcu, 2001). It is usually necessary to modify the optimal candidate sentence or to piece together partially-matched fragments if the domain is unrestricted, due to sparse data issues.

It is natural in our system to index the translation table using some form of the interlingua. The complexity of the index determines the degree of correspondence between the matched translation pairs: a more detailed index potentially leads to a closer match, but at the increased risk of search failures due to data sparseness. We use very lean semantic information, encoded as key-value pairs, to index our automatically generated translation corpus. The KV string is derived from the semantic frame by the GENESIS system using trivial generation rules. We can further reduce data sparseness by masking values of certain keys (e.g., city names, months and dates, etc.) during the retrieval stage, and re-inserting them in the surface string. This is equivalent to the technique of replacing lexical entries with classes during example-based matching described in (Brown, 1999).

Given the thin KV index, it is possible to have sentences with very different syntactic structure to map to the same index, as shown by the examples in Table 1. Any Chinese sentence asking about the temperature in Boston can be mapped to these three example sentences (plus many other possibilities), and vice versa. This could become a useful feature for language learning: we can present multiple translation choices to a student for increased variability.

English:	The temperature in Boston. Tell me the temperature in Boston. What is the temperature in Boston?
KV:	WEATHER: temperature CITY: Boston

Table 1: Three English sentences and their corresponding key-value string.

Figure 3 summarizes the example-based translation process.

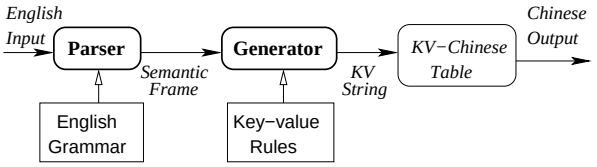


Figure 3: Schematic diagram for example-based translation method. Note: KV = Key Value.

5 Evaluation

We evaluated our system using English speech data, recorded from phone calls to the publicly available JUPITER weather information system (Zue et al., 2000). Our test data consists of 695 utterances selected from a set of held-out data. Utterances whose manually-derived transcription can not be parsed by the English grammar are excluded from the evaluation, since they are likely to be out-of-domain sentences and would simply contribute to null outputs. The test data have on average 6.5 words per utterance. The recognizer achieved 6.9% word error rate and 19.0% sentence error rate on this set.

A bilingual judge rated the translation quality based on grammaticality and fidelity, where in input sentence is either a manual transcription or a recognizer output. Performance differences between these two modes reflect degradations caused by speech recognition errors.

We adopt the strategy of preferring the generation-based output if it can be accepted by the Chinese grammar. The generation method is able to achieve high fidelity in the translation, preserving syntactic correspondence between English and Chinese as much as possible. We back off to the example-based method if the generation method failed. Table 2 summarizes the number of translations produced by each method. The "yield" on ASR

outputs is lower, due to parsing and KV generation failures on severely corrupted ASR outputs.

Category	Manual	ASR
Generated	606	592
Example	59	48
Failed	30	55

Table 2: Number of translations in each category (generated, by example, or failed both) for manual transcriptions and automatic speech recognition (ASR) outputs on a set of 695 utterances.

Table 3 summarizes the subjective ratings of the translation outputs. The translations are likely to be well-formed for both manual and ASR transcriptions, due to the "acceptance" check imposed by the Chinese grammar. However, accuracy on ASR outputs is expectedly lower, due to mis-recognized semantic entities, such as city names. This is unlikely to be a serious issue for the language students, because our system echos a paraphrase of the recognized input to keep the user informed during the interaction. A closer look at the system outputs also revealed that minor syntactic errors in ASR outputs seldom cause translation degradation, due to the example-based mechanism. As long as a robust parse can be found containing all the semantically important fragments, we are able to produce an appropriate translation using the lookup mechanism. The overall translation accuracy is 94.3% (including "perfect" and "acceptable") for manual transcriptions, and 90.2% for ASR outputs.

Quality	Manual	ASR
Perfect	613	577
Acceptable	43	50
Wrong	9	13

Table 3: Number of "perfect," "acceptable," and "wrong" translations, for manual transcriptions and automatic speech recognition (ASR) outputs.

6 Future Work

While we believe that our methodology will be effective for language learning applications, we have yet to demonstrate that this is the case. We also want to port it to many other applications besides weather, to form a suite of lesson plans on different topics. Our syntax-based formulation of the English grammar will ease the burden of porting to other domains. We are also conducting research on automatic techniques to induce the Mandarin grammar, given the English-to-Mandarin translation framework.

Acknowledgments Support for this research was provided in part by the Cambridge/MIT Initiative.

References

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

R. D. Brown. 1999. Adding linguistic knowledge to a lexical example-based translation system. In *Proc. TMI*, Chester, England.

M. Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proc. ACL*, Madrid.

B. Cowan. 2004. PLUTO: A preprocessor for multilingual spoken language generation. Master's thesis, MIT, Cambridge, MA.

R. E. Frederking, A. W. Black, R. D. Brown, A. Rudnicky, J. Moody, and E. Steinbrecher. 2002. Speech translation on a tight budget without enough data. In *Proc. Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, Pennsylvania.

Y. Gao, B. Zhou, Z. Diao, J. Sorensen, H. Erdogan, and R. Sarikaya. 2002. A trainable approach for multilingual speech-to-speech translation system. In *Proc. HLT*, San Diego, CA.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL*, Edmonton, Canada.

L. Levin, A. Lavie, M. Woszczyna, and A. Waibel. 2000. The Janus III translation system. *Machine Translation Journal*, 15(1-2). Special Issue on Spoken Language Translation.

D. Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proc. ACL*, Toulouse, France.

M. Rayner and P. Bouillon. 2002. A flexible speech to speech phrasebook translator. In *Proc. Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, Pennsylvania.

M. Rayner and S. Carter. 1997. Hybrid processing in the spoken language translator. In *Proc. ICASSP*, Munich, Germany.

S. Sato. 1992. CTM: an example-based translation aid system using the character-based match retrieval method. In *Proc. COLING*, Nantes, France.

S. Seneff and J. Polifroni. 2000. Dialogue management in the MERCURY flight reservation system. In *Proc. ANLP-NAACL, Satellite Workshop*, Seattle, WA.

S. Seneff, C. Wang, and T. J. Hazen. 2003. Automatic induction of n-gram language models from a natural language grammar. In *Proc. Eurospeech*, Geneva.

S. Seneff, C. Wang, and J. Zhang. 2004. Spoken conversational interaction for language learning. In *These proceedings*.

S. Seneff. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1).

M. Tang, X. Luo, and S. Roukos. 2002. Active learning for statistical natural language parsing. In *Proc. ACL*, Philadelphia.

T. Veale and A. Way. 1997. Gaijin: A template-driven bootstrapping approach to example-based machine translation. In *Proc. NMNLP*, Sofia, Bulgaria.

V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington. 2000. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1).