High-quality Speech Translation in the Flight Domain

Chao Wang and Stephanie Seneff

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

Abstract

Portability is an important issue to the viability of a domain-specific translation approach. This paper describes an English to Chinese translation system for flight-domain queries, utilizing an interlingua translation framework that has been successfully applied in the weather domain. Portability of various components is tested, and new technologies to handle parse ambiguities and ill-formed inputs are developed to enhance the translation framework. Evaluation of translation quality is conducted manually on a set of 432 unseen flight-domain utterances, which are translated into Chinese using a formal method and a new robust back-off method in tandem. We achieved 96.7% sentence accuracy with a rejection rate of 7.6% on manual transcripts, and 89.1% accuracy with an 8.6% rejection rate on speech input. A game for language learning using the translation capability is currently under development.
Index Terms: speech translation, domain portability, natural language understanding, natural language generation.

1. Introduction

For the past several years, we have been developing spoken dialogue systems to help a student of a foreign language acquire proficiency, by allowing them to engage in spoken conversation with the computer in the new language [1]. In order for such a system to be effective, it must also be able to provide assistance at any time by acting as an always-present tutor. The main task of the software tutor is to provide translation assistance, either from the native language into the second language to teach the student how to formulate a query in the new language, or from the second language into the native language to help the student understand the system response in the new language.

Previously, we have developed a high-quality speech translation system in the context of a weather domain [2]. The system adopted an interlingua framework, using formal rules for parsing and generation. A target language grammar was applied to validate the translation outputs by verifying that they could parse. If the parse failed, then an example-based translation mechanism was invoked as a back-off. In this case, semantic information encoded as [key: value] pairs was used to retrieve a suitable candidate from a pre-compiled corpus of translation examples. While our approach is domain-specific, we emphasize portability in both the parsing and generation components, so that translation capabilities in other domains can be quickly developed using the same technology.

In this paper, we present our work on developing translation capability for queries in the flight information domain. The level of complexity of the flight domain queries is substantially higher than that of queries typical of the weather domain, so translation of these queries, which must be of extremely high quality in order not to mislead the student, is a very challenging task. In making our natural language understanding component more portable, we have adopted a largely syntactic grammar for parsing. While adapting the grammar for a particular domain becomes very straightforward, the syntactic approach has the disadvantage of increased ambiguities in parse theories. The so-called "PP-attachment" problem [3] is much more prevalent in the flight domain than in the weather domain, where the queries contain a much richer set of prepositional phrase modifiers. In addition to using a probability model in the parser, we implement an effective rule-based mechanism to rearrange parse output to reduce attachment errors.

The formal *parse-generate* method is capable of producing very high-quality translation when the inputs are within the coverage of the rules. However, it is fragile on "novel" inputs, when a user uses expressions unforeseen by the rule developer, or when the inputs are corrupted by recognition errors. Previously, we have developed an example-based translation (EBT) method as a back-off for improved robustness. For the flight domain, we experiment with a different mechanism which also backs off to using [key: value] information as an interlingua. While the two methods are somewhat equivalent in essence, we believe the new method is better at dealing with data sparseness issue, and more elegant in its implementation.

Our approach is similar to [4] in that translation is modelled as a cascade process of parsing and generation. However, [4] uses only semantic information in the interlingua, and adopts a statistical approach for parsing and generation. In contrast, out interlingua representation captures both syntactic and semantic information, and the parsing and generation adopt a rule-based framework. We also share many parallelisms with [5], which also focused on the travel domain. One eminent similarity is that both have been working towards the goal of a more general grammar framework.

In the remainder of the paper, we will first give an overview of our translation system. We then describe how we addressed two challenges we encountered in the flight domain, namely, parsing ambiguities and ill-formed inputs. Evaluation of English to Chinese translation quality is reported in Section 5, followed by conclusions and future work.

2. Overview

Figure 1 shows a diagram of the translation procedure, for the scenario of English to Chinese translation. The procedure begins with parsing the input string and deriving an interlingua representation of the input, which we term "semantic frame."[1] We use the

[1]This nomenclature is somewhat misleading as the semantic frame encodes both syntactic and semantic information.

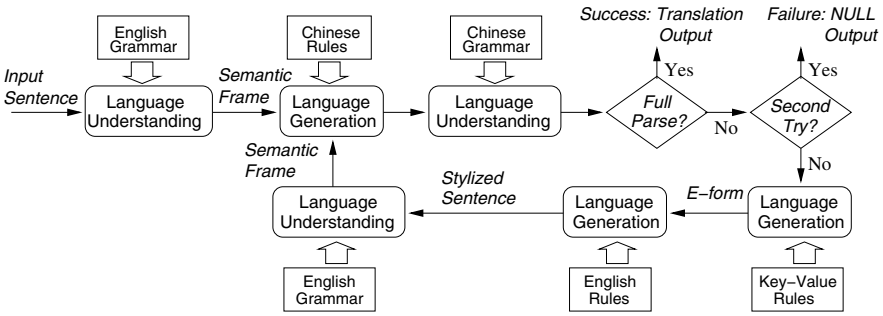

Figure 1: *Diagram of translation procedure involving a formal parse-generate method and a robust back-off method using [key: value] information.*

```
{c wh_question
   :trace "what" :auxil "link"
   :topic {q flight :quant "def" :dem "next"
       :pred {p leave
           :topic {q city :name "new york"}}
       :pred {p destination
           :topic {q city :name "ithaca"}}}}
```

Figure 2: *Example semantic frame for the sentence "What is the next flight leaving New York to Ithaca?" Some details irrelevant to Chinese generation have been omitted in the Figure.*

TINA [6] system, which utilizes a context-free grammar to define allowable patterns, augmented with a probability model to select among ambiguous parse theories. In making our natural language understanding component more portable, we have adopted an approach that uses mainly syntactic information in the majority of the parse tree rules. Semantics are introduced near the terminals, mainly involving adjectives, verbs, nouns and proper noun classes. Rules for general semantic concepts such as dates and times are organized into sub-grammars that are easily embedded into any domain. The flight domain and weather domain thus share a common core grammar encoding the syntactic structure typical of database query applications. Figure 2 shows an example semantic frame resulting from the parsing step.

To generate well-formed Chinese strings from the semantic frame, we use the GENESIS generation system [7], which works from a lexicon of context-dependent word-sense surface strings for each vocabulary item, along with a set of recursive rules to specify the ordering of constituents in the generated string. It supports sharing of generation rules among a large number of elements that follow a common generation pattern. Typically, there is a default rule for each of the three major constituent types: "clause," "topic," and "predicate," that covers the most common generation pattern for each type. Most clause templates can be shared by different domains: they are mainly governed by the syntactic structure in the semantic frame and are hence domain-independent. Templates for certain generic predicates (e.g., "want," "tell," "show," etc.) and topics (e.g., "pronoun") can also be reused. Domain-dependency comes mainly from domain-dependent topics and predicates. To the extent that they can use the generic template, the main task for a developer is to figure out the ordering of the constituents in the templates, i.e., the position of predicates relative to the noun (before or after the noun), and the relative ordering of predicates (e.g. "flight leaving New York to Ithaca" vs. "flight to Ithaca leaving New York"). Similar to the understanding component, we have also isolated the generation rules for time expressions into a separate file, so that they can readily be reused by different domains.

The Chinese string from the parse-generate step is further processed through the Chinese understanding system, which also adopts a largely syntactic grammar for improved domain portability. If it fails to parse, a more robust but less precise back-off method is adopted, based on a simple [key: value] representation of the concepts encapsulated in an electronic form (e-form). This representation is derived automatically through simple generation rules from the semantic frame. A stylized English sentence is then generated from the e-form specification, which is essentially a simplified paraphrase of the original input (see Section 4 and Figure 4). This simplified English paraphrase then goes through the parse-generate channel again to produce a Chinese translation, which becomes the translation output if it can be parsed by the Chinese understanding system. If the translation generated from the e-form path fails to parse in the Chinese grammar, the system apologizes for being unable to translate the user's query. This technique assures that, if the user can accurately imitate the provided translations, the Chinese grammar will be able to process their query. There is of course a risk of rejecting a perfect translation due to gaps in coverage of the Chinese grammar, but over time these gaps will eventually be filled.

3. Handling Parse Ambiguities

A disadvantage of a strictly syntactic approach to parsing is that there are typically many more cases of alternative parse theories than in a less generalizable grammar with strong ties to semantic relationships. A serious issue is the so-called "PP-attachment" problem [3], which is widespread, for instance, in the pattern `vbo object pp`, where the prepositional-phrase (`pp`) can attach to the verb phrase (`vbo`) or to the noun phrase representing its direct object. Consider the sentence, "Show me flights leaving in the morning *at 10 a.m.*." The phrase *"at 10 a.m."* is three-ways ambiguous, being interpretable as "leave at 10 a.m.", "flight at 10 a.m." or even "show me at 10 a.m." Incorrect attachment can be undetectable if the paraphrase is into a language with the same overall word-sequence structure as English. But for paraphrases from English into Chinese, an erroneous attachment will often lead to an erroneous paraphrase.

The probability model utilized by our parsing engine [6] can learn to disambiguate with high accuracy based on the preferred probability solution, because it captures dependencies beyond context-free rules by conditioning on the external left-context parse category when predicting the first child of each parent node. However, this assumes that it has been provided with a large corpus of correctly parsed training utterances. We do not have a corpus of bracketed flight domain data, and acquiring such a corpus would be very labor intensive. Instead, we found that an effec-

```
{c rule
   :in ( {p see} {p show} ... )
   :contains ( {p temporal} ... )
   :to {q flight} }
```

Figure 3: *Example rewrite rule specifying a reattachment of a temporal modifier, to be repositioned from modifying the main verb to instead attach to any noun phrase containing "flight" as its main noun.*

tive and simple solution is to carefully arrange the training data such that it first parses all the sentences that contain a single noun phrase (there are many, due to the fact that this is spontaneous conversational speech). A further benefit can be gained by arranging both the noun phrases and the more complex sentences in order of increasing string length, in an attempt to train it first on simple patterns that don't contain ambiguities. Since the system updates its probability model incrementally, it learns which prepositional phrases are observed frequently in the noun phrases before encountering the more complex sentence structures.

While this strategy greatly reduces the ambiguity problem, it does not totally eliminate it, and so we sought a further solution to repair the remaining errors. The semantic frame, derived automatically from the parse tree, encodes the PP-attachment decision in its hierarchy. One possible solution is to develop a separate statistical model capturing semantic relationships in the semantic frame. An *N*-best list of parse candidates can then be rescored via this semantics-driven probability model for possible reranking. We explored this idea using a number of different schemes for training the semantics, but we were ultimately unsatisfied with the outcome.

In fact, given that the attachment of the prepositional phrases in the flight domain are highly unambiguous from a pragmatic point of view, it turned out that a more effective solution was to simply empower developers to write rules to capture such domain knowledge. It is after all improbable that, in "I am looking for a flight on United with a stop in Denver," *on United* would attach to "look," or *with a stop in Denver* would attach to either "look" or "United." We use reconstruction rules which specify semantic frame patterns that would trigger a surgical step to rearrange problematic constituents in specific ways. An example of such a "rearrangement rule" is given in Figure 3. It required very little developer effort to specify the set of rewrite patterns that were needed to repair the remaining errors observed in parsing our flight domain corpus. These rules are of course domain-dependent, and would have to be respecified by developers for each new domain. However, they seem to offer a viable solution to the PP-attachment problem for restricted domains.

4. Translating Robust Parse Sentences

We applied our parse-generate technique to a set of some 25,000 flight domain English queries obtained from prior data collection efforts via the telephone-based MERCURY system [8]. We found that most of the challenging problems due to differences in the two languages could be handled through the use of existing mechanisms in our GENESIS language generation system [7]. A novelty of the approach we are using in the flight domain as compared with our previous research in the weather domain is a different strategy for handling the back-off generation from the e-form. In the weather domain, we adopted an example-based translation approach [9], which involved finding a matching template from a translation corpus given an e-form specification. To cope with

Original Sentence	⇒	Paraphrase
Could you tell me which ones would be evening flights that would leave around 7pm.	⇒	Which ones are flights in the evening leaving around 7pm.
What airline is the flight originating in Atlanta on November 7th at noon and arriving at San Francisco at 2:10pm.	⇒	What is the airline for the flight from Atlanta to San Francisco departing at noon on November 7th arriving at 2:10pm.
Tell me about flights leaving from Atlanta and going to Charlotte North Carolina next Monday I need to know about flights that arrive Charlotte between 4:15 and 5:30pm.	⇒	I would like flights from Atlanta to Charlotte North Carolina departing next Monday arriving between 4:15 and 5:30pm.

Figure 4: *Examples of robust parse sentences and their stylized paraphrases.*

sparse data problem, values of certain attributes (such as dates and city names) were masked by the corresponding class name during retrieval, and reinserted in the matching template's surface string. Since the flight domain is far more complex than the weather domain in terms of both the number of possible attributes and the complexity of the clause structure, there turns out to be a much greater sparse data problem as well as a more challenging task of variable substitution. We therefore decided, instead, to utilize our generation system to supply direct generation rules from the e-form into a simplified English paraphrase. We then apply the standard formal method to the simplified string. Examples of English queries alongside their simplified English paraphrases are shown in Figure 4. As illustrated by the examples, the stylized paraphrase has a fairly simple sentential structure, with most of the attributes being attached to the main noun phrase in a specified order. This ensures that the resulting paraphrase has a straightforward translation and will be more likely to succeed through the normal translation channel.

The back-off translation mechanism is incorporated into the full translation system as depicted by the diagram in Figure 1.

5. Evaluation

The MERCURY corpus has over 25,000 transcribed utterances in the flight domain, which also includes a test set that have been held out from acoustic/language model training to facilitate speech recognition evaluations. Although our rule-based translation framework does not rely heavily on data for statistical training (with the exception of the parsing grammar as discussed previously), the data are very useful at helping us quickly identify gaps in the translation rules: we feed the training data through the translation system, and focus on sentences which failed to produce a translation output after the parsibility check in the target language. The process is iterated until a reasonable percentage of sentences can be successfully translated.

To evaluate the speech translation system, we selected a subset of test utterances from the held-out data which have not been used in developing the translation rules. We excluded from the original pool any meta-queries, such as "repeat," "start over," or "good bye," which are commands directed to the system. We also precluded any utterances that were only one word long (e.g., "yes," "no"), since the translation of these sentences is uninteresting. Finally, we excluded utterances whose transcript failed to parse (13.6% of the original pool of data), since our methods provide no

	No. Utts	Percentage
1) Total Utts	432	
2) Recognizer output parsed	426	98.6% of 1)
3) Direct translation generated	426	100% of 2)
4) Direct translation accepted	384	90.1% of 3)
5) KV translation generated	33	78.6% of 2)-4)
6) KV translation accepted	11	33.4% of 5)

Table 1: *Breakdown of percentage of 432 test utterances processed to different stages in the translation procedure shown in Figure 1. (Note: KV = [key: value])*

Mode	P+A	I	F	Yield	Accuracy
text	356+30	13	33	92.4%	96.7%
speech (*1*-best)	314+31	43	44	89.8%	88.9%
speech (*10*-best)	317+35	43	37	91.4%	89.1%

Table 2: *Manual ratings of translation quality on a set of 432 unseen test utterances with text and speech inputs. (Note: P = Perfect, A = Acceptable, I = Incorrect, F = Failed,* $Yield = \frac{P+A+I}{P+A+I+F}$, $Accuracy = \frac{P+A}{P+A+I}$*)*

means for translating them. Our test set consists of 432 utterances, with an average size of 5.6 words per utterance. Speech recognition makes use of the SUMMIT landmark-based recognizer [10], which achieved a 10.6% word error rate on these data.

Three system configurations were evaluated: 1) with the manual orthographic transcription as input, 2) with the recognizer outputting only the top hypothesis, and 3) with the recognizer producing an *N*-best list, from which the parser selects the highest scoring candidate, considering both acoustic and linguistic scores. The translation quality was rated by a bilingual speaker of Chinese and English, following the same procedures outlined in [9]. Three categories were used in the manual rating: `Perfect (P)`, `Acceptable (A)`, and `Incorrect (I)`. A fourth category `Failed (F)` is used to represent NULL translation outputs.

Table 1 shows the percentages of the utterances that made it to various stages of the translation procedure depicted in Figure 1, for speech input with the recognizer producing *10*-best hypotheses. Nearly all of the utterances (98.6%) were able to obtain at least one parsed hypothesis from the *10*-best list and produce a direct translation string. Less than 10% of the generated translations failed to parse in the Chinese grammar, over a quarter of which (78.6% x 33.4%) were recovered through the robust back-off mechanism by using [key: value] information. We observed that some of the failures were due to gaps in the Chinese grammar. Overall, 91.4% of the original utterances produced a parsable Chinese translation.

Table 2 summarizes the rating results of translations produced with the three input modes described previously. Our translation system achieved good translation performance in text mode: it produced parsable Chinese translations for over 92.4% of test utterances, 96.7% of which were either perfect or acceptable. This verifies that the formal method is capable of very high-quality translation when the inputs are within the coverage of the rules. The yield in speech mode did not change by much, although accuracy dropped by about 8% relative to that of the text mode. The increased error rate is largely due to speech recognition errors affecting content words, e.g., misrecognized dates, times, or cities. Since our system provides a paraphrase of the recognized string to the user, hopefully this type of error would not mislead the language learner. The *10*-best mode produced more translations than the *1*-best mode with roughly the same level of accuracy.

6. Conclusions and Future Work

We developed an English-to-Chinese speech translation system for flight domain queries. Evaluated on an unseen test set of 432 utterances, the system produced perfect or acceptable translations for over 80% of the total data, while rejecting about 8.6% inputs. Our approach to portability is also proven to be effective.

The translation capability is used as the core component of a translation game system currently under development for language learning. The game can be played via the Web, currently configured to assist a native English speaker practicing Chinese. The student would be given a randomly generated English utterance, and would be tasked with providing an equivalent Chinese utterance, either by speaking or by typing in pinyin. The system would compare the utterance's meaning with one obtained from its own internal translation, and if the utterance is judged correct, would congratulate the student and provide a new utterance to translate. As the game progresses, the system would keep track of how many turns it took the student to successfully translate each utterance, and would gradually increase or decrease the difficulty level depending on the student's performance. Explicit user enrollment would allow the system to personalize the difficulty level to match the student's capabilities.

While we have developed high-quality translation capabilities in two domains (weather and flight information), it remains to be seen if our approach is scalable to larger domains. We are currently investigating the scalability issue within the context of translating general travel-related phrases.

7. Acknowledgements

This work was supported in part by the Cambridge MIT Institute and by ITRI Research Labs.

8. References

[1] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in *Proc. of InSTIL*, Venice, Italy, 2004.

[2] C. Wang and S. Seneff, "High-quality speech translation for language learning," in *Proc. of InSTIL*, Venice, Italy, 2004.

[3] D. Hindle and M. Rooth, "Structual ambiguity and lexical relations," *Computational Linguistics*, vol. 19, no. 1, pp. 103–120, 1993.

[4] Y. Gao, B. Zhou, Z. Diao, J. Sorensen, and M. Picheny, "MARS: A statistical semantic parsing and generation-based multilingual automatic translation system," *Machine Translation*, vol. 17, pp. 185–212, 2002.

[5] L. Levin, A. Lavie, M. Woszczyna, and A. Waibel, "The Janus III translation system," *Machine Translation*, vol. 15, no. 1-2, 2000, Special Issue on Spoken Language Translation.

[6] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, 1992.

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

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

[9] C. Wang and S. Seneff, "High-quality speech-to-speech translation for computer-aided language learning," *ACM Transactions on Speech and Language Processing*, 2006, accepted for publication.

[10] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.