INTERLINGUA-BASED TRANSLATION FOR LANGUAGE LEARNING SYSTEMS

John Lee, Stephanie Seneff

Spoken Language Systems Group
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139 USA
{jsylee,seneff}@csail.mit.edu

ABSTRACT

This paper concerns our recent research in developing high-quality spoken language translation for restricted domains. The intended application is a spoken-language translation aid for a student of a foreign language. A significant novelty of the work is in leveraging an existing English-to-Mandarin translation system in the weather domain both to provide a corpus of sentence pairs for training and to induce an initial version of the parsing grammar for translation in the reverse direction. Using an interlingual approach, we are able to reject strings that fail to parse, yielding high accuracy on any translations provided to the student. On a test set of 369 naturally spoken Mandarin queries, the translation was judged incorrect for fewer than 3% of the query transcripts. A statistical phrase-based translation system performed significantly worse, when trained on the same material.

1. INTRODUCTION

For the past several years, we have been developing systems that will enable a student of a foreign language to practice conversation in a non-threatening environment [1, 2]. This research has built upon our previous work in developing multilingual dialogue systems [3, 4, 5]. We now have in place a system intended to help a native English (L1) speaker learn Mandarin (L2), by engaging him/her in conversation in the weather domain. The student converses over the telephone, and can speak in either English or Mandarin at any time. Mandarin queries about the weather are answered in Mandarin, whereas English queries are automatically translated into Mandarin. The student can then attempt to imitate the resulting Mandarin query to push the conversation forward.

Our interest here is in creating a reversed scenario in which a native Mandarin speaker is learning English. Ideally, reversing the language roles would simply require reassigning L1 and L2 in a top-level system control file. However, the reversed system is missing a critical component: a high-quality translation capability from Mandarin to English. While we already have a capability to *understand* Mandarin queries, the grammar for understanding only needed to capture the semantic content of the utterance and hence did not preserve sufficient detail for accurate translation.

Section 2 of this paper introduces the two methodologies we explored for translation (interlingual and statistical), while Section 3 elaborates on the interlingual methodology in particular. Section 4 presents evaluation results on both automatically generated queries and naturally spoken Mandarin queries. The final section provides a discussion of future plans.

2. MACHINE TRANSLATION OVERVIEW

Machine translation systems are employed in various types of applications, in which users have rather different expectations of their translation capabilities. These differences have implications on both the translation methodology and the performance evaluation criteria.

2.1. Translation Applications

For *information access* applications, such as translating foreign web pages, the system should provide wide coverage and be robust on ill-formed input, since the user wants to obtain as much information as possible from unrestricted text. Grammatical correctness, while desirable, is not critical, so long as the essential information is conveyed.

Further, the user is assumed to be proficient in the output language, and so there need not be any restrictions on the linguistic variety of the translated sentence. In most cases, there are multiple valid translations for a single sentence. The BLEU score [6], which considers multiple reference translations, is well suited to these applications and has become the standard evaluation metric.

For *second-language learning* applications, the expectations are almost complementary. The system must be able to provide grammatical, near-perfect translation. Wide coverage is not required, since the typical conversation revolves

0-7803-9479-8/05/$20.00 © 2005 IEEE

around particular domains of interest. In fact, if a sentence is out-of-domain or ill-formed, it would be better for the system not to attempt translation, than to risk misinforming the student.

Moreover, since the user is a learner of the output language, the translation should exhibit a consistent style, vocabulary set and grammatical constructions, focused on the lesson being taught. For instance, in a lesson on the future tense, it would not be desirable to translate an example sentence in the present tense, even if both tenses happened to be acceptable. When multiple translations are acceptable, there is usually one (the "canonical" sentence type) that a language teacher would consider to be most relevant with respect to the pedagogical purpose at hand. Hence, one would like the translation to be not only correct, but also as close in syntax and vocabulary to this reference translation as possible. Thus, in the first experiment, we have adopted an extremely strict evaluation criterion, the word error rate (WER), borrowing from metrics typically used for speech recognition tasks. In the second experiment, where reference translations were not available, we performed a manual evaluation.

2.2. Translation Methodologies

For *information access* applications, the phrase-based statistical machine translation method has been shown to be most effective (e.g., [7]). The translation model does not require any linguistic knowledge beyond pairs of aligned sentences, for which the alignments can be automatically learned using general-purpose alignment tools such as [8]. As a result, they require little manual effort, and easily scale up in coverage, whenever large bilingual corpora are available. The lack of linguistic knowledge, however, limits their ability to model long-distance dependencies, such as tense usage.

For *second-language learning* applications, we argue that the interlingual method (e.g., [1], [9] and [10]) is more suitable. This method performs syntactic and semantic analysis on an input sentence, and maps it to a hierarchical meaning representation called the "semantic frame". Rich linguistic knowledge, encoding long-distance dependencies, is easily incorporated into the semantic frame, and is taken into account during generation. The design of such a meaning representation is challenging for a wide domain, but has been shown to be feasible for restricted ones. A further advantage of this method, not yet exploited in this work, is the relative ease with which feedback can be presented to the student. The semantic frame, as a canonical meaning representation, is useful for detecting inappropriate or missing linguistic features. The corresponding grammatical error can then be explained to the student.

3. INTERLINGUAL FRAMEWORK

This section briefly outlines the translation framework for our interlingual method. Full details can be found in [1]. Our framework requires three components for L1-to-L2 translation, as shown in Figure 1: a *natural language understanding (NLU) system* [11], which maps a sentence in L1 to a semantic frame encoding syntax and semantics, a *transfer* phase, which modifies the semantic frame to account for linguistic properties unique to L2, and a *natural language generation (NLG) system* [12], which produces a well-formed surface string in L2.

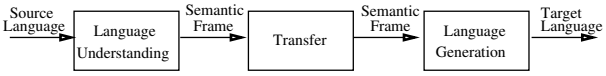

Fig. 1. *Schematic diagram of our translation framework.*

In our NLU system, a set of context-free rules describes the sentence structure. The grammars that are designed for our translation system typically incorporate both syntactic and semantic information simultaneously. At the higher levels of the parse tree, major syntactic constituents, such as subject, predicate, direct object, etc., are explicitly represented through syntax-oriented grammar rules. The syntactic structures tend to be domain-independent, capturing general syntactic constraints of the language. Near the leaves of the parse tree, major semantic classes, such as *weather* and *date*, are constructed according to semantics-oriented grammar rules. The semantic structures tend to be domain-dependent, capturing specific meaning interpretations in a particular domain. The semantic frame is derived directly from the resulting parse tree through a simple set of rules to assign parse tree categories to syntactic roles.

3.1. Grammar Development

We utilized a corpus of over 45,000 transcribed English utterances taken from the JUPITER weather information domain [3], of which 691 were held out for testing. This domain allows the public to ask about weather over the phone. The average utterance length is 6.0 words.

A preliminary NLU grammar for Mandarin was induced semi-automatically from an English grammar utilizing a recently developed grammar induction algorithm. This involved parsing the English utterances and paraphrasing them into Mandarin with a pre-existing English-to-Mandarin translation system [1], while automatically aligning the Mandarin paraphrase with the original English parse tree. A rearranged parse tree could then be automatically transformed into Mandarin grammar rules. For details of this process, please see [13]. An important consequence is that the induced Mandarin grammar yields a semantic frame that closely parallels the original English-derived semantic

frame. The induced grammar was manually altered, particularly in the higher level nodes of the parse tree, to better reflect certain characteristics of Mandarin syntax. A set of constraints was also supplied to support the characteristic movement of temporals and locatives to the head of the sentence in Mandarin.

3.2. Transfer Stage

A semantic frame is, in theory, a language-independent representation of an input sentence. In practice, however, it is often influenced by the linguistic properties of the input language. Consider the semantic frames in Figures 2 and 3, obtained from the English sentence, "*Will there be any snow in the Midwest tomorrow?*" and its Mandarin paraphrase, respectively.

Just as Mandarin lacks a number of linguistic features, such as auxiliaries, determiners, and tense, so does the frame in Figure 3. This impoverished frame would have generated the ungrammatical English surface string "*Midwest have snow tomorrow?*". This problem is reflected in the errors made by the induced Mandarin grammar in [13], the majority of which were deletions of English function words that had no counterparts in Mandarin.

```
{verify :rhet "there" :aux "link" :tense "future"
   :topic {weather_act :name "snow"
              :pred {temporal
                        :topic {day :name "tomorrrow"} }}
              :pred {locative
                        :topic {region
                                  :quant "def" :name "midwest"}
}}}
```

Fig. 2. *Semantic frame for the sentence, "will there be any snow in the Midwest tomorrow?"*

```
{verify :pred { have
   :topic {weather_act :name "snow"
              :pred {temporal
                        :topic {day :name "tomorrrow"}}}
              :pred {locative
                        :topic {region
                                  :name "midwest"}
}}}}
```

Fig. 3. *Semantic frame for the sentence, "ming2tian1 zhong1xi1bu4 you3 mei2 you3 xue3," a Mandarin paraphrase of the English sentence in Figure 2. The Mandarin literally reads, "tomorrow Midwest have not have snow".*

The transfer stage [14] reinstates such missing features by manipulating the semantic frame based on specific requirements of the output language. In the case of English, for example, a :topic under a **have** predicate is pulled to the top level of the frame. Appropriate auxiliaries and tense properties are also inserted based on formal rules. The frame in Figure 3 is thus mapped to the one in Figure 2.

3.3. Generation Stage

Our NLG system maps a semantic frame to a surface string using formal generation rules. These rules specify the order in which components in the frame are to be processed into substrings, and consult a generation lexicon to obtain multiple word-sense surface-form mappings and appropriate inflectional endings.

4. EVALUATION EXPERIMENTS

We evaluated the translation quality of both the interlingual and statistical methods, for two different data sets: automatically generated Mandarin queries, and naturally spoken Mandarin queries collected from native Mandarin speakers conversing with our MUXING weather domain.

4.1. Experiment 1: Automatically Generated Queries

4.1.1. Bilingual Corpus Creation

Figure 4 shows a block diagram of our training and testing procedure. A new Mandarin-English bilingual corpus was created from the English JUPITER corpus as follows. The output of a pre-existing English-to-Mandarin (E:M) translation system [1] was used as the Mandarin reference string. An English-to-English (E:E) paraphrase output, rather than the original English utterance, was used as the English reference string. This is because the E:E paraphrases, produced by the NLG system, tend to exhibit a simpler and more consistent style, and hence serve better in a language learning application than the original utterances.

Thus, we are interested in seeing to what extent the indirect path English → Mandarin → English differs from the direct paraphrase path, English → English. Figure 5 shows three examples of queries processed through all these steps to produce our training (E:M and E:E) and testing (M:E and E:E) data. An example alignment between a reference and hypothesis string is shown in Figure 6.

The quality of this automatically generated bilingual corpus was further assessed with the Mandarin grammar developed in §3.1. The Mandarin translations (E:M) that could not be fully parsed by the grammar were excluded from further evaluation. Out of the original JUPITER English test set, about 8% of the Mandarin translations failed to parse. Roughly half of these contained grammatical errors introduced by the E:M translation system, most frequently due

[1] In a previous experiment, 94.3% of the Mandarin translations of this system were judged *Perfect* or *Acceptable* [1].

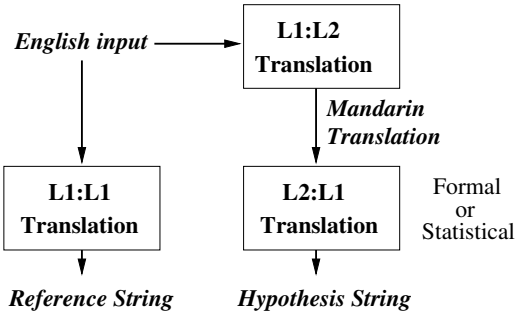

Fig. 4. *Illustration of procedure for training and testing. The statistical method is trained on a corpus of utterance pairs of the "Mandarin Translation" and the "Reference String" based on 45,000 user queries. The English translation from the Mandarin string (for both the interlingual and statistical methods) is compared with the English-to-English "translation" of the original string. Evaluation uses the standard speech recognition criterion of substitution/insertion/deletion error rates computed from aligned reference and hypothesis strings.*

to incorrect word order. One quarter were out-of-domain, ill-formed or with false starts. The other quarter (2% of the full set) were legitimate Mandarin translations, which our Mandarin grammar was unable to parse because of out-of-vocabulary items, or lack of coverage in the context-free rules. The filtered test set contained 633 Mandarin sentences.

4.1.2. Statistical Method

To compare our interlingual translation method with a statistical method, we utilized PHARAOH [15], a state-of-the-art phrase-based translation system. It was trained and tested on the same bilingual corpus as described in §4.1.1.

4.1.3. Results

The interlingual method outperformed the statistical one by 3.4% in terms of absolute word error rate, as shown in Table 1. The best performing system, however, was obtained by cascading the interlingual system with a statistical English-to-English "translation" system.

The English-to-English statistical system was trained as follows. Its source sentences were the English translations of the interlingual system on the Mandarin strings in the bilingual corpus ("hypothesis string" in Figure 4); some of these sentences contained Mandarin-influenced grammatical errors. Its target sentences were the reference English paraphrases ("reference string" in Figure 4). When testing, the output of the interlingual system was fed to this English-to-English statistical system. This cascaded approach achieved only 10% word error rate.

What is the temperature in Dallas?	E-in
What is the temperature in Dallas?	**E:E**
da2la1si1 de5 qi4wen1 shi4 duo1shao3	E:M
(Dallas of temperature is how much?)	Literal
What is the temperature in Dallas?	**M:E**
Will it be windy in Boston on friday?	E-in
Will it be windy in Boston friday?	**E:E**
xing1qi1 wu3 bo1shi4dun4 feng1 hui4 da4 ma5	E:M
(Friday Boston wind will big q-particle?)	Literal
will it be windy in Boston friday?	**M:E**
What places do you know in China?	E-in
What places in China do you know?	**E:E**
ni3 zhi1dao4 zhong1guo2 de5 shen2me5 di4fang4?	E:M
(You know China of what place?)	Literal
What places in China do you know?	**M:E**

Fig. 5. *Three examples of queries processed through the series of steps used in our training and evaluation methodologies. The statistical translation system is trained on the E:M and E:E corpus pairs, and both systems are tested on aligned string error rates for the M:E and E:E pairs. NOTE: E = English; M = Mandarin.*

System	Subs.	Del.	Ins.	Total
STATISTICAL	4.8%	4.5%	5.3%	14.7%
INTERLINGUAL	4.5%	5.7%	1.2%	11.3%
CASCADED	4.0%	3.3%	2.8%	10.0%

Table 1. *Word error rates in Experiment 1.*

For both systems, the majority of the errors were caused by translation variants of Mandarin words that did not significantly alter the meaning of the sentence. The questions "*what is the weather like in Boston*," "*what is the weather for Boston*," and "*how is the weather in Boston*" are all paired with the same Mandarin string in our bilingual corpus. Similarly, "*know / know about*," "*warnings / advisories*" are equivalent expressions in most contexts in our domain. The performance gains of the cascaded system over the interlingual method alone were mostly due to reductions in these errors.

Lack of context in short sentences led to other errors. For example, both "*In Boston*" and "*Boston*" are acceptable translations to the Mandarin "*bo1shi4dun4*" ("*Boston*"), depending on the previous sentence in the dialogue. Without context, the interlingual method favors the singular "*advisory*" over the plural "*warnings*," which in many cases turned out to be the original word.

Some of the errors produced by the statistical method were significant for language learning applications. At times, it inserted a spurious "*there*," likely due to an incorrect phrase alignment. More consistently, in sentences with future temporal expressions, it preferred "*is*" instead of the correct "*will*," a long-distance dependency that it was unable to learn. In a sentence like "*ming2tian1 zhong1xi1bu4*

REF: what	IS	the chance	THAT IT WILL	rain	****	in Boston tonight?
HYP: what	WILL	the chance	**** OF ****	rain	BE	in Boston tonight?

Fig. 6. *Example alignment of a reference string with a hypothesis string that realizes a 55% (5/11) error rate according to our criterion, in spite of being a perfectly fine translation.*

you3 mei2 you3 xue3" (see Figure 3), both "*is*" and "*will*" are possible translations of "*you3*," whose tense is not systematically marked. Its proper translation thus depends on temporal expressions elsewhere in the sentence.

The statistical method was unable to handle some other long-distance grammar constraints. Consider the question "*ni3 zhi1 bu4 zhi1dao4 ming2tian1 hui4 xia4yu3 ma5*" ("*do you know if it will rain tomorrow*"), which was translated as "*do you know about will it rain tomorrow*". While the two halves of the sentence are perfectly good phrases on their own, they form an ungrammatical sentence when combined.

4.2. Experiment 2: Naturally Spoken Queries

4.2.1. Bilingual Corpus Expansion

The bilingual corpus in §4.1.1 was enhanced with a set of 847 Mandarin utterances, obtained from user interactions with our MUXING weather system [4]. Compared to the utterances used in the first experiment, these have a substantial number of new Mandarin geographical terms, a few novel grammatical constructions, and more varied and colloquial styles. Their transcripts were split into two halves for training (TRANS_{train}) and testing (TRANS_{test}).

The transcripts in the training set were analyzed to extend the coverage of our Mandarin grammar. The grammar was then used to filter TRANS_{train} and TRANS_{test} in the same way as in the previous experiment: those utterances whose transcripts could not be fully parsed were eliminated from further consideration. For TRANS_{test}, 87% of the Mandarin sentences were parseable. 20% of the unparsed sentences had grammatical mistakes or false starts, 32% had out-of-domain words, mostly geographical or weather terms, and 48% had previously unseen expressions for queries and greetings. The filtered TRANS_{test} consisted of 369 sentences.

Since reference English translations were not available, the improved interlingual system was utilized to generate English translations for sentences in the TRANS_{train} set. The resulting Mandarin-English pairs were then *added* to the bilingual corpus described in §4.1.1 for training the statistical system. While no evaluation was carried out on these English translations, they should be no worse than the translations of the test set, of which fewer than 3% were judged incorrect (see §4.2.3).

To study how much translation quality degrades with respect to speech recognizer errors, we considered the output of a speech recognizer based on the SUMMIT [16] framework. The recognizer was used in the MUXING system and performed at 13% word error rate. This additional test set, SPEECH_{test}, consists of the ASR output corresponding to the TRANS_{test} set.

4.2.2. Evaluation Criterion

We examined the English output sentences manually, and rated their quality, given the original Mandarin sentence as a reference, as *Perfect*, *Acceptable*, or *Wrong*. A sentence is *Acceptable* if it conveys the correct information, but has minor syntactic or tense mistakes, or untranslated proper names. It is considered *Wrong* if any pertinent information is missing or mistranslated.

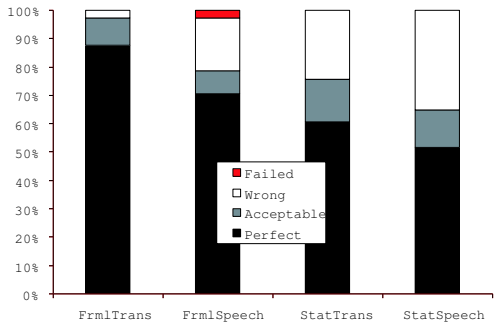

Fig. 7. *Results for manual evaluations of translations in Experiment 2. Note: "Frml" = Interlingual method; "Stat" = Statistical Method; "Trans" =* TRANS_{test}*; "Speech" =* SPEECH_{test}.

4.2.3. Results

The average rankings of the English output sentences are shown in Figure 7. As expected, translation quality degraded when speech recognizer outputs were used instead of transcripts.

In terms of the proportion of *Perfect* sentences, the interlingual method outperformed the statistical one by over 16% in both the transcribed and automatically recognized test sets. On the transcribed set, fewer than 3% of the translations of the interlingual system were judged incorrect. This number increased to 19% when translation was attempted on recognizer outputs. Parse failure on these out-

puts, all but three of which were ungrammatical, contributed 2.7% to the incorrect translations.

In some cases, the statistical method suffered from data sparseness in the alignment process. One example is the phrase "*wen4 yi1 xia4*," a polite expression for "*ask*," which appeared only once in the training set. In the output sentence, the word "*wen4*" was not translated and "*yi1 xia4*" was incorrectly translated.

Lastly, we also examined the utterances that were filtered out of TRANS$_{test}$ due to parse failure, but were grammatically correct and in-domain (48% of the failed subset; 6% of the total: see §4.2.1). Although not part of the evaluation shown in the figure, these utterances could still yield a translation using the statistical method. However, only one of the resulting English output sentences was judged *Perfect*, and two *Acceptable*. Such parse failures thus served as an effective technique to flag sentences whose translations are likely to be erroneous, thus avoiding misinforming the student.

5. CONCLUSION AND FUTURE PLANS

We have presented an interlingua-based translation method for language learning systems. Leveraging an existing English-to-Mandarin translation system in the weather domain, a translation system in the reverse direction was semi-automatically created and evaluated. On both automatically generated and naturally spoken queries, this system compares favorably with a statistical approach.

Although the statistical approach does not perform as well, it provides a fast, automated alternative to build a reversed translation system, training on a bilingual corpus generated by the existing English-to-Mandarin system.

In the future, we plan to apply the techniques described here to generate high-quality Mandarin-to-English translations in a number of different domains for which we have available English corpora. This will allow us to expand the range of conversational topics for students to practice communication in a foreign language. We have already begun an effort in the flight domain, for which we can make use of tens of thousands of transcribed English utterances.

We would also like to adapt our translation framework to the task of grammar correction, e.g., correcting errorful English sentences of non-native English speakers. We envision this task as an "*errorful English*"-to-"*good English*" translation, where the transfer stage is responsible for detecting and correcting grammar mistakes.

6. ACKNOWLEDGEMENTS

The authors would like to thank Chao Wang for her help both in running the statistical translation experiments and in providing a tool to expedite the manual evaluation process. This research was supported in part by a research grant from the Defense Language Institute, mediated through MIT Lincoln Laboratory, and in part by the Cambridge MIT Institute.

7. REFERENCES

[1] C. Wang and S. Seneff, "High-quality Speech Translation for Language Learning," in *Proc. InSTIL*, Venice, Italy, 2004.

[2] S. Seneff, C. Wang, M. Peabody, and V. Zue, "Second Language Acquisition through Human Computer Dialogue," in *Proc. ISCSLP*, Hong Kong, 2004.

[3] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A Telephone-based Conversational Interface for Weather Information," *IEEE Trans., Speech and Audio Processing*, vol. 8(1), pp. 85–96, 2000.

[4] C. Wang, S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue, "MUXING: A Telephone-access Mandarin Conversational System," in *Proc. ICSLP*, Beijing, China, 2000.

[5] V. Zue, S. Seneff, J. Polifroni, M. Nakano, Y. Minami, T. J. Hazen, and J. Glass, "From JUPITER to MOKUSEI: Multilingual Conversational Systems in the Weather Domain," in *Proc. MSC*, Kyoto, Japan, 2000.

[6] K. Papineni, S. Roukos, T. Ward, and W. Zhu "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proc. ACL*, Philadelphia, PA, 2002.

[7] P. Koehn, F. Och, and D. Marcu, "Statistical Phrase-based Translation," in *Proc. HLT-NAACL*, Edmonton, Canada, 2003.

[8] F. Och and H. Ney, "Improved Statistical Alignment Models," in *Proc. ACL*, Hong Kong, China, 2000.

[9] Y. Gao, B. Zhou, Z. Diao, J. Sorensen, and M. Picheny, "MARS: A Statistical Semantic Parsing and Generation-based Multilingual Automatic Translation System," in *Machine Translation*, vol. 17, pp. 185–212, 2002.

[10] K. Precoda, H. Franco, A. Dost, M. Frandsen, J. Fry, A. Kathol, C. Richey, S. Riehemann, D. Vergyri, J. Zheng, "Limited-Domain Speech-to-Speech Translation between English and Pashto," in *HLT/NAACL* demo sessions, Boston, MA, 2004.

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

[12] L. Baptist and S. Seneff, "GENESIS-II: A Versatile System for Language Generation in Conversational System Applications," in *Proc. ICSLP*, Beijing, China, 2000.

[13] J. Lee and S. Seneff, "Translingual Grammar Induction," in *Proc. Interspeech*, Jeju Island, Korea, 2004.

[14] B. Cowan, "PLUTO: A Preprocessor for Multilingual Spoken Language Generation," M.S. thesis, MIT, Cambridge, MA, 2004.

[15] P. Koehn, "PHARAOH: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models," in *Proc. AMTA*, Washington DC, 2004.

[16] J. Glass, "A Probabilistic Framework for Segment-based Speech Recognition," in *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.