Two-Stage Translation: A Combined Linguistic and Statistical Machine Translation Framework

Yushi Xu, Stephanie Seneff
Spoken Language Systems Group, Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{yushixu, seneff}@csail.mit.edu

Abstract

We propose a two-stage system for spoken language machine translation. In the first stage, the source sentence is parsed and paraphrased into an intermediate language which retains the words in the source language but follows the word order of the target language as much as feasible. This stage is mostly linguistic. In the second stage, a statistical MT is performed to translate the intermediate language into the target language. For the task of English-to-Mandarin translation, we achieved a 2.5 increase in BLEU score and a 45% decrease in GIZA-Alignment Crossover, on IWSLT-06 data. In a human evaluation of the sentences that differed, the two-stage system was preferred three times as often as the baseline.

1 Introduction

There are two main approaches for machine translation nowadays: statistical and linguistic. Statistical machine translation, which has recently become the dominant paradigm, provides a powerful ability to learn lexical mappings and translation models without extensive manual effort. However, such a system requires a large amount of training data in order to learn meaningful models. This is always a problem for domains with only a limited corpus. Another defect is that, even with enough data, statistical MT is poor at handling long distance reordering. The fact that it does not know anything about the underlying syntactic structure of the sentence limits its algorithm to only penalize reordering according to distance constraints, a severe limitation. On the other hand, a linguistic approach does not require much data, but needs considerably more human expertise to design the rules. Linguistic methods are good at capturing and making use of the overall structure of the sentence, and their behavior tends to be better understood, and thus explainable by humans. But they are challenged to perform as well as statistical methods in areas like word sense selection.

In order to produce high quality spoken language translation, where the source- and target-languages differ dramatically, and only a small amount of training data is available, we propose a combined MT system which takes advantage of the strong points of both linguistic and statistical MT systems. We call our method two-stage translation, in which the structural construction is explicitly done by a linguistic language generation method, and the word-sense disambiguation and lexical mapping is done by a statistical system. We experiment with the approach in an English-to-Chinese translation scenario, in which the source and target languages differ significantly. We apply the method to IWSLT-06 data, drawn from a domain of spoken conversational sentences collected in the travel domain. This corpus is small (~40K sentences), and has relatively broad topics, including weather, food, shopping, city navigation, passport control, time scheduling, etc., which would require a great deal of human effort to come up with accurate translation rules in purely linguistic MT systems. The corpus also differs significantly from news corpora, which are used to train many statis-

tical MT systems. The sentences contain a large percentage of wh-questions, and the corpus is much smaller than typical news corpora, which will cause the statistical MT system to face severe sparse data problems.

2 Previous Research

There are a few systems that use purely linguistic machine translation. Examples include the dialogue interaction and translation assistance systems in the weather domain [Zue et al., 2000] and the flight domain [Seneff and Polifroni, 2000]. Wang and Seneff have shown that a formal parse-generate method is capable of producing high quality translations when the input is within a limited domain [Wang and Seneff, 2006]. Their system uses a parser to process the input sentence into a meaning representation, and then a rule-based language generator to generate the target sentence string. The same paradigm applies to the translation game system in the hobby and schedule domain [Chao et al., 2007]. These systems produce highly accurate translations, but they only work on very limited domains.

A typical example of a purely statistical MT system is the phrase-based statistical machine translation system MOSES [Koehn et al., 2007]. It learns a phrase table as well as a translation model from the training corpus. This system has been used in many tasks, but usually it requires millions of training utterances from a parallel corpus to train a good model.

Besides the above two types of MT systems, a number of researchers have been trying to incorporate syntax information into statistical systems. [Yamada and Knight, 2001] performed *tree-to-string* translation to directly transform the source-language parse tree into the target-language string. [Zhang et al., 2007] modeled statistical phrase reordering based on the source-language parse tree. [Wang et al., 2007] reordered the tree nodes using hand-coded linguistic rules. [Zhang, Zen and Rey, 2007] applied shallow parsing on the source side, and automatically extracted chunk reordering rules based on word alignment. [Habash, 2007] adopted a similar idea, but used source-language dependencies to extract the reordering rules. Another system that adopts an approach that is similar to ours is the system by [Simard et al, 2007]. They first translated the sentences using rule-based MT, then used statistical MT as an automatic post-editing layer. All this research has shown that a hybrid system can produce better results.

In terms of domain and language, many of the previous work has been applied to written corpora. Only a few have focused on the spoken domain [Shen et al, 2007; Carpuat and Wu, 2007]. Although many researchers have been working on translation from Chinese into English, far less research has been devoted to the reverse direction.

3 Two-Stage Translation

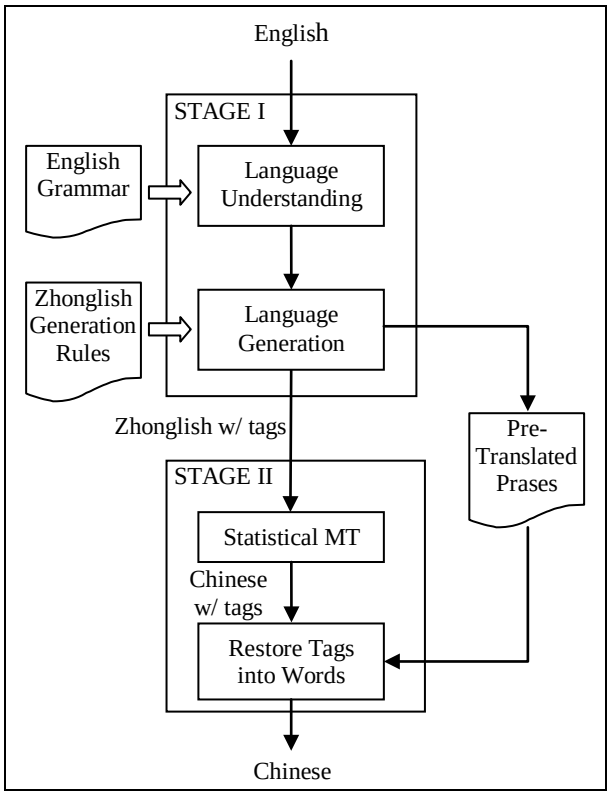

Figure 1. Framework of the system

Figure 1 shows the framework of our system. We separate the translation process into two stages. In the first stage, the linguistic method is used to perform structural reconstruction. The source sentence, which in our experiment is in English, is parsed into a meaning representation by a language understanding component. The meaning representation serves as the input to the language generation component. The output of Stage I is a string in an intermediate language, which we call "Zhonglish". The Zhonglish output string maintains most of the words in English, but adopts Chinese word order and Chinese-unique structure. This stage is

different from conventional reordering in that it is generation-based. English-unique words like "the" may be deleted, and Chinese-unique words like "BA"(把) and "MA"(吗) may be inserted. Some simple phrases are fully translated into Chinese in this stage, and are replaced with special tags, such as $date, or $time. In the second stage, a Zhonglish-Chinese statistical MT is performed. Since the Zhonglish sentences already have the Chinese structure, little reordering is required in this stage. The main function of the statistical system is word-sense disambiguation and lexical mapping. Finally, we restore the tagged phrases into actual Chinese phrases.

3.1 Stage I: English-Zhonglish Translation

3.1.1 Language Understanding

In the language understanding step, the input sentence is parsed and converted into a "linguistic frame," a hierarchical meaning representation that encodes both syntactic structure and semantic knowledge but discards temporal order. We use the TINA [Seneff, 1992] system, which utilizes a context-free grammar to define allowable patterns, augmented with a probability model to help select among ambiguous parses, and a feature passing mechanism to deal with movement and agreement constraints. After parsing, a separate step converts the parse tree into a linguistic frame. This is accomplished by visiting all nodes in the parse tree in a top-down left-to-right path, and building up the linguistic frame incrementally by consulting a simple mapping table. An example of the output linguistic frame, for the sentence, "Where did you put the book I bought yesterday?," is shown in Figure 2.

Besides the context-free grammar rules, the TINA system contains a trigram spatial-temporal probability model which is trained automatically. It also has a powerful trace mechanism, which moves the words in the surface form back to their deep syntax structure position. For example, in Figure 2, the word "where" is restored to a position under the verb phrase "put" and the word "book" is restored to its position as the object of the verb "buy." This mechanism is especially useful and important when generating into languages which have a very different set of rules for movement, such as Chinese, as we will illustrate in the next section.

```
{c wh_question
  :auxil "xdo"
  :topic {q pronoun
          :name "you"
          :number "pl" }
  :mode "past"
  :pred {p put
         :topic {q object
                 :clause_object {c noun_clause
                                 :topic {q pronoun
                                         :name "i"
                                         :number "first" }
                                 :pred {p buy
                                        :topic {q trace_object
                                                :quantifier "def"
                                                :noun "book" }
                                        :mode "past"
                                        :pred {p temporal
                                               :topic {q rel_date
                                                       :name "yesterday" }}}}}
         :pred {p trace_pred
                :trace "where" } } }
```

Figure 2. Linguistic frame for input sentence "Where did you put the book I bought yesterday."

3.1.2 Language Generation

We use the GENESIS [Baptist and Seneff, 2000] system for language generation, together with its preprocessor PLUTO [Cowan, 2004]. The preprocessor fills missing features that are specific to the target language. Both the main processor and the preprocess operate with rule templates to generate a surface string from the linguistic frame. The rules dictate the order in which the constituents are to be generated. The three basic constituents in a linguistic frame, namely, clauses, topics (noun phrases), and predicates (broadly including verb phrases, prepositional phrases, and adjective phrases) each have a default generation rule template, which handles the majority of the constituents. In addition, every specific constituent can have its own unique generation rule, or can be a member of a special class that share a unique generation rule. Reordering of the constituents can easily be done by manipulating the generation rules. The syntax of the rules also allows pull/push actions which can handle cross-level movement, such as the wh-movement in English. The GENESIS system includes a context-dependent lexicon, which can disambiguate the word-sense according to the part-of-speech and various flags set during the generation procedure.

Previously, we had developed a set of English generation rules, which can paraphrase the linguis-

tic frames back into English with high quality. We use this as a reference and a starting point from which to develop the rules for Zhonglish, through the following steps.

Step 1, Undo Overt Movement. Since Chinese usually retains the moved constituents of English wh-movement rules in their deep structure position, it was relatively straightforward to *disable* the generation rules that restored the surface form English string for wh-marked NP's.

Step 2, Omit Inflection and Do-Support. Chinese is a morphology-impoverished language. Inflections of verbs and nouns are thus omitted. And the auxiliary "do" is deleted.

Step 3, Add Chinese Ordering Rules. Chinese basically shares the same SVO structure as English, but differs in a couple of ways. Modifiers like temporals, prepositional phrases and relative clauses are usually preposed. There are also some special constructions in Chinese. For example, the BA-construction rule, which will be discussed in more detail below, realizes a clause in SOV order instead of SVO order. The Zhonglish generation rules order the words according to Chinese word order as much as possible, and also insert Chinese function words that do not exist in English, like "BA" and "DE."

Input: where did you put the book I bought yesterday?
Reference: 你 把 我 昨天 买 的 书 放 哪里 了 ?

Step 1: did you put I bought the book yesterday where?
- 你 放 我 买 - 书 昨天 哪里
Step 2: you put I buy the book yesterday where?
你 放 我 买 - 书 昨天 哪里
Step 3: you BA I yesterday buy DE the book put where?
你 把 我 昨天 买 的 - 书 放 哪里

Figure 3. Example outputs after each of three steps of conversion from English to Zhonglish generation languages.

Figure 3 illustrates how the above three steps affect the output string for the linguistic frame in Figure 2. Ideally, if we can produce a perfect Zhonglish sentence, we can simply apply statistical lexical mapping rules to produce a perfect Chinese sentence. But this is not always possible, since translation is usually not simply a literal word-to-word conversion.

3.1.3 Dealing with Random Constructions

Our purpose is to separate the translation into two stages, so that the linguistic stage can exercise control over the sentence structure, and as a consequence, the statistical MT system will have less work in figuring out the reordering. Thus, in order to train the statistical component to a high-quality lexicon, phrase table and translation model, it is essential to generate the intermediate language in the way the parallel target sentence is expressed. However, languages have many optional constructions, which complicate the process of deciding how to order constituents. For example, in both English and Chinese, the ordering of multiple prepositional phrases modifying a noun or verb is often ambiguous. One especially challenging example is the Chinese BA-Construction rule, which is often optional. Figure 4 shows how a source English sentence can be translated into Chinese equally well with and without BA-Construction. If, in the training data, the parallel target sentence is (a), we don't want the language generation system to output Zhonglish with the word order of (b).

Input: Please open your mouth.
Translation (a): 请 张开 你 的 嘴
Please open your mouth
Translation (b): 请 把 你 的 嘴 张开
Please BA your mouth open

Figure 4. Alternative translations of an English sentence with and without BA-Construction.

We deal with this problem by two means. First, the conditions that trigger BA-Construction are carefully examined in our training data. In the situation where BA-Construction is obligatory, the output sentence will always use BA-Construction. When it's optional, the decision is made by taking advantage of the presence of the parallel sentence during training. When producing the linguistic frame, we examine the parallel sentence to find certain features, such as the presence of "BA", and add this information into the linguistic frame generated from the English sentence. Instead of parsing the Chinese sentence, the feature is captured by simple string comparison. This may introduce false features, but they will be ruled out by the constraints written in the generation rules. Only when the feature is set *and* all the constraints are satisfied, the language generation system will produce a sentence with the construction.

3.1.4 Pre-Translating Phrases

Not all the lexical items can be better translated by a statistical system than a linguistic system. Numbers, times and dates are good examples. Because the training data can never cover all of the possible numbers or dates, the results from a statistical MT system are often remarkably wrong for these kinds of translations, as illustrated in Figure 5.

Input: eight singles and eight quarters, please.
Output: 一 美元 和 八 八 个 两 角 五分 的 辅币。
(one dollar and eighty eight quarters)
Input: my watch loses ten minutes a day.
Output: 我 的 表 一 天 慢 三 十 分钟。
(my watch loses thirty minutes a day)

Figure 5. Problematic translation of numbers for statistical MT system.

In our approach, the linguistic system's rules fully translate some short phrases and replace them with a unique tag before entering Stage II.

The pre-translation is done directly by the language generation component. The GENESIS system includes a context-dependent lexicon. Instead of generating "eight," for example, we use the lexicon to map "eight" directly to the Chinese character "八". Then we replace the actual translation with sequential abstract tags, such as $number1, $clock_time1, etc. The pre-translated words and the corresponding abstract tag names are stored in a mapping table for later access.

Approximately the same process is applied to the parallel Chinese sentence in training, except that, when processing Chinese, we use a fragment parser to tag the corpus. The fragment parser only looks for phrases that satisfy some specific grammar, e.g. numbers, times and dates, and ignores everything else. The fragment parser is very robust in that it won't produce parse failures for complex sentences.

After this process, the sentences are "abstract" to some extent. We generate a language model using the abstracted sentences, which will be used in Stage II. Finally, the outputs produced by the statistical MT system are post-processed to replace the unique tags with their corresponding Chinese word strings.

Although statistical MT can perform a similar pre-translation, we do this explicitly for two reasons. First, the language model generated by the "abstracted" sentences is stronger. Second, although we currently only apply pre-translation on numbers, dates and times, this is an extendable method. We can gradually extend this capability to pre-translate more phrases that the linguistic system is confident of.

3.1.5 Dealing with Ambiguous Parses

Parsing ambiguity is a traditional problem for linguistic systems. If the wrong English parse is chosen, good generation rules will generate bad Zhonglish sentences. The most significant example is ambiguous PP-attachment problem.

In Figure 6, the sentence "May I see the jacket in the window?" has two ambiguous parses. The prepositional phrase "in the window" can be attached to either the verb "see" or the noun "jacket." The two different attachments result in two different Zhonglish outputs. "In the window" ends up landing right before the word it is attached to. Also, when the PP is modifying a noun phrase, an additional function word "DE" is necessary. In order to generate a correct Zhonglish sentence, in this case, the second choice, we need a way to choose the right parse.

Input: may I see the jacket in the window?
Reference: 我 能 看看 橱窗 里 的 夹克衫 吗

Parse 1: [may I [see [the jacket] [in the window]]?]
Zhonglish: I may in the window see the jacket ma?
我 能 在 橱窗 里 看看 夹克衫 吗

Parse 2: [may I [see [[the jacket] [in the window]]]?]
Zhonglish: I may see the in the window DE jacket ma?
我 能 看看 在 橱窗 里 的 夹克衫 吗

Figure 6. Ambiguous parses of a sentence exhibiting the PP-attachment problem.

One way to solve this problem, of course, is to improve the parser. The parser does have a probability model that can bias it towards the correct answer, but sparse data problems lead to inevitable errors. However, the explicit striking differences in the surface forms when paraphrased into Zhonglish imply that we can use statistical language modeling techniques applied to the Zhonglish string to potentially reduce errors.

Figure 7 gives an example of the idea. We write two sets of Zhonglish generation rules. One set is the ordinary Zhonglish generation rules. The other one we name as "conservative PP-reordering rules," in which only those PPs that are *not* ambi-

guous are reordered to the preposed position. The possibly ambiguous PPs, typically in the situation of VP NP PP, are still left behind as if expressed in English. So the Zhonglish sentences generated from this set of rules are guaranteed to have all *preposed* PPs in correct positions. We train an *n*-gram language model on the output of this first run.

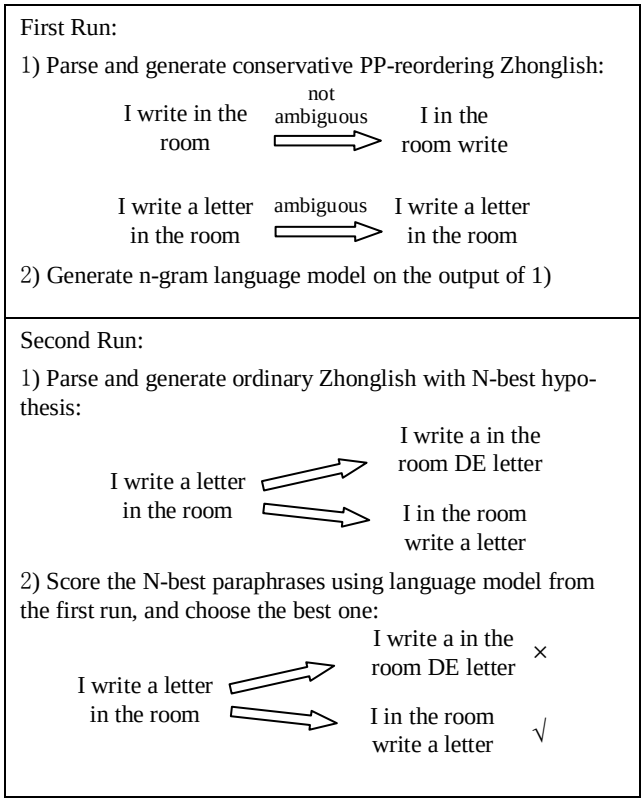

Figure 7. Example of disambiguation process.

Then in a second pass, we use the ordinary Zhonglish generation rules, in which, regardless of possible ambiguity, all the PPs are reordered into the *preposed* positions. We let the language understanding system output N-best linguistic frames, and let the language generation system generate N-best Zhonglish sentences for the linguistic frames. Finally, the *n*-gram language model (trained using the above procedure) scores the N-best list and selects the best-scoring one. Because the generation rules force all the PPs to be preposed this time (except in some rare cases where the prepositional phrase is a complement rather than an adjunct), the statistics of the postposed PP in the language model will never be used. Only the *n*-grams relating to the preposed PPs will take effect. They contain information about whether the PP prefers to be attached to the specific verb or to the specific noun. Thus, compared to the raw result coming out from the parser, the one chosen by the *n*-gram language model will have a much greater likelihood to be correct. Figure 8 shows some examples of results before and after this process.

Input	may I see the jacket in the window?
Before	[may I [see [the jacket] [in the window]]?] I may in the window see the jacket ma?
After	[may I [see [[the jacket] [in the window]]]?] I may see the in the window DE the jacket ma?
Input	I leave my money in the safety box in my room.
Before	[I [leave [[my money] [in the safety box]] [in my room]].] I BA my in the safety box DE money leave in my room.
After	[I [leave [my money] [[in the safety box] [in my room]]].] I BA my money leave in the in my room DE safety box.

Figure 8. Examples of best parses and their Zhonglish paraphrases before and after the PP-attachment disambiguation process.

3.2 Stage II: Zhonglish-Chinese Translation

The Zhonglish-Chinese translation is done by a standard phrase-based statistical MT system. Ideally, the Zhonglish and Chinese pair only differs in the language of the words, and reordering is not at all necessary at this stage. But we still turn on the reordering in the statistical MT system to account for the cases that the linguistic system missed.

4 Experiment and Evaluations

4.1 Experimental Setup

The statistical MT system we use is MOSES. This also serves as our baseline. The maximum reordering distance is set to 6. The corpus is the IWSLT-06 data, which is a domain of travel and tourist information. The training set consists of 39,952 parallel sentences, among which about 20% are wh-questions. We use two held-out development sets: dev-set 1 as our development set to do minimal error training, and dev-set 2 as our test set. Both sets have approximately 500 sentences. The experiment is English-Chinese translation.

4.2 Results

The baseline system is trained with the full training set. However, for our system, the language understanding system cannot parse all of the source sentences. We throw away the sentences that cannot be parsed or can only produce a partial parse result. This gives us a parsable set of 35,672 parallel sentences for training, which is about 89% of the full training set. The same thing is done to the development set. The BLEU scores reported in Table 1 are all tested on the parsable set of the test data, which consists of 453 sentences.

	Baseline (trained with full set)	Baseline (trained with parsable set)	Our Approach (trained with parsable set)
BLEU	31.48	30.78	33.33

Table 1. BLEU score of baseline and our system.

As shown in Table 1, even with 11% less training data, our approach realized a 1.85 point improvement on BLEU score over the baseline. When the baseline is restricted to train on only the parsable data, our approach gained over 2.5 BLEU points. Figure 9 shows some comparisons of results between the baseline and our approach.

a) what time does the bus for boston leave?
B: 什么 时候 的 巴士 从 波士顿 出发 ？
(The bus of what time leaves from Boston?)
O: 这 趟 去 波士顿 的 巴士 什么时候 出发 ？
(When does this bus for Boston leave?)

b) that comes to five dollars and thirty-two cents.
B: 总共 两 美元 三 十 五 美分 。
(Altogether two dollars thirty-five cents.)
O: 总共 是 五 美元 三 十 二 美分 。
(Altogether is five dollars thirty-two cents.)

c) it's cherry blossom season.
B: 这 是 cherry blossom 季节 。
(This is cherry blossom season.)
O: 它 是 樱花 的 季节 。
(It is cherry blossom's season.)

Figure 9. Different translation results from baseline system and our system. B denotes baseline, and O denotes our system. Literal translations of the Chinese sentences are shown in brackets.

In analyzing the results, we concluded that the gain is mostly due to three reasons. First, proper structural construction helps the output translation to be in the correct word order. Example a) in Figure 9 exhibits this point. The output from the baseline is almost a straight word-to-word translation. The meaning is quite different from the original English sentence. But with our approach, the sentence is correctly translated. Secondly, pre-translation helps. Our system can always get the numbers, times and dates correct, as shown in example b). Finally, example c) shows that, with the two stages, the statistical MT system can better align the parallel sentences, thus producing a lexicon and phrase table with higher quality.

Despite the improvement we got for BLEU score, we observed that this is not a very reliable measurement for our experiment. Chinese has freer word order and more optional words than English, but we only have one reference translation for each test sentence. Furthermore, some reference sentences have typos and some are not judged as the natural way of expressing the meaning by native speakers. Therefore, we also did a human evaluation on the subset of the test set where the two systems' outputs differ. The human judge is given the original English sentence and is asked to judge which translation is better, or that they are equally good (equally bad). Table 2 shows the result. It is very clear that our approach did much better than the baseline statistical MT. The results in both lines are statistically significantly different ($p < 0.001$).

	# baseline better	# our system better
Baseline with full training set	33	99
Baseline with parsable training set	26	98

Table 2. Human evaluation of the baseline and our system.

4.3 GIZA Alignment Crossover

The BLEU score and the human evaluation showed the overall performance of the system. We also want to test how well Stage I does. Since the output of Stage I is a manmade language, which doesn't have any ground-truth to compare with, we developed another way to evaluate: the GIZA alignment crossovers. Perfect Zhonglish would achieve an ideal goal where, when a Zhonglish sentence is aligned with the corresponding Chinese sentence, all the word pairs are in sequence, i.e. the crossover is zero. So, in Stage II, the lower the number of crossovers on the training data, the better Stage I performs.

Figure 10 exemplifies how we actually count the crossover. Each cross is counted as one. If an alignment crosses two other alignments, it is counted as two. This means long distance disorder is worse than local disorder, which conforms to human perception.

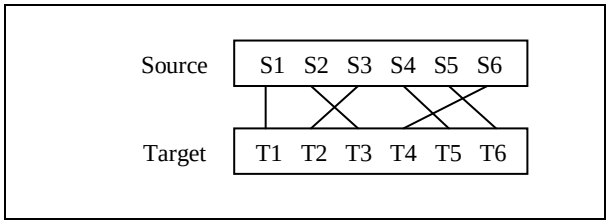

Figure 10. Example of calculating GIZA-alignment crossover. In this example, the crossover equals to 3.

	Avg. Crossover Per Sentence	Normalized Avg Crossover	% of Zero-Crossover Sentences
E-C Pair (full training set)	11.3	0.97	22.0
E-C Pair (parsable training set)	10.52	0.95	22.5
Z-C Pair (parsable training set)	5.91	0.52	40.5

Table 3. GIZA-Alignment Crossovers for original English-Chinese Pair and our Zhonglish-Chinese Pair. The normalized average crossover is obtained by normalizing on the length of the source sentences.

This measurement is objective and easy to obtain. It is very intuitive from the number to see how well our Stage I is doing. Table 3 lists the average crossover of the training data of the baseline system (pair of English and Chinese sentences) and our system (pair of Zhonglish and Chinese sentences).

Our Zhonglish-Chinese pair has only about 55% as many crossovers as the original English-Chinese pair. And we have 18% *absolute* more sentences that are crossover-free. This is a substantial improvement. We also measure how our dealing with random constructions and ambiguous parses affect the crossover. Table 4 gives the result.

The figures show positive results of our methods for handling the two problems. Especially for the handling of ambiguous parses, the decrease in crossover indicates that more PP's are in the correct position; i.e., they are attached to the correct phrases.

	Avg. Crossover Per Sentence	Normalized Avg Crossover	% of Zero-Crossover Sentences
Z-C Pair	6.08	0.54	39.4
Z-C Pair (+RC)	6.03	0.53	40.0
Z-C Pair (+RC, +AP)	5.91	0.52	40.5

Table 4. Effects of random construction and ambiguous parse handling. RC stands for Random Construction. AP stands for Ambiguous Parses.

5 Conclusion and Future Work

We presented a two-stage approach for machine translation. In the first stage, the translation from English input to Zhonglish output involves extensive linguistic knowledge about both the English grammar (to parse the English sentence) and the Chinese grammar (to generate the Zhonglish sentence). During the Zhonglish generation process, we take advantage of the parallel corpus to provide additional information from the parallel sentence and to overcome the problem of random constructions. We also pre-translate some numbers, times and dates during generation. The ambiguous PP-attachment problem is handled with an *n*-gram language model generated by conservative PP-reordering rules. We have verified both by the conventional BLEU score metric and human evaluation that our method outperforms the purely statistical MT system. The BLEU score increases by 2.5 percent, and a human evaluator prefers the two-stage system 3:1 over the baseline. We also measured the performance of the linguistic stage by GIZA alignment crossovers. After reconstructing the sentence structure by linguistic methods, the crossover decreases by 45 percent absolute.

Within the linguistic stage, we throw away 11% of the data due to the limited ability of the language understanding system. However, we saw that the baseline system can gain an improvement of about 0.7 BLEU points by using this part of the data. This implies that, if our system can somehow use the unparsable data, we may realize additional improvement. One way is to simply improve the coverage of the parser through further grammar rules, and another is to allow partial parses. We can break one unparsable sentence into pieces and process each piece, then glue the resulting Zhonglish pieces together. But this may cause problems when a reordering needs to cross the boundary of the two pieces. Another possible way is to simply

use the unparsable English sentences and pretend they are Zhonglish. This would likely inject noise into the system that could weaken the performance of the well-modeled training data, but both ways are worth trying.

Another possible space to explore is linguistic pre-translation. Currently we are only pre-translating numbers, times and dates. But we can also pre-translate other phrases that we are confident to translate linguistically. Gradually, as more and more phrases are pre-translated, the whole system will move towards the linguistic side. And with more words replaced by the corresponding tags, the sparse data problem becomes less severe, and very likely, the statistical stage will perform better.

Finally, we also plan to apply this approach to other machine translation task such as English-Arabic and Chinese-English. We are confident that this approach can have positive results in other tasks as well.

Acknowledgments

This research has been supported by the Delta Electronics Environmental and Educational Foundation. We would like to thank Chao Wang for her advice and guidance in this research.

References

Yasuhiro Akiba, Eiichiro Sumita, Hiromi Nakaiwa, Seiichi Yamamoto, Hiroshi G. Okuno. 2003. Experimental Comparison of MT Evaluation Methods: RED vs. BLEU. In *Proc. of MT Summit IX, New Orleans, USA, 2003*.

Lauren Baptist and Stephanie Seneff. 2000. Genesis-II: A Versatile System for Language Generation in Conversational System Applications. In *Proc. of ICSLP, Beijing, China, 2000*.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proc. of EMNLP, Prague, Czech Republic, 2007*.

Chi-Hyu Chao, Stephanie Seneff, and Chao Wang. An Interactive Interpretation Game for Learning Chines. *Proc. of the Speech and Language Technology in Education (SLaTE) Workshop, Farmington, Pennsylvania, 2007.*

Boxing Chen, Jun Sun, Hongfei Jiang, Min Zhang and Ai Ti Aw. 2007. I2R Chinese-English Translation System for IWSLT 2007. In *Proc. of IWSLT, Trento, Italy, 2007*.

Brooke A. Cowan. 2004. PLUTO: A Preprocessor for Multilingual Spoken Language Generation. Master's Thesis, MIT, Cambridge, Massachusetts, 2004.

Nizar Habash. 2007. Syntactic Preprocessing for Statistical Machine Translation. In *Proc. of MT Summit XI, Copenhagen, Denmark, 2007.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *ACL demonstration session, Prague, Czech Republic, June 2007*.

Daniel Marcu, Wei Wang, Abdessamad Echihabi and Kevin Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proc.of EMNLP, Sydney, Australia, 2006.*

Stephanie Seneff. 1992. TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics, Vol. 18, No. 1, 1992*.

Stephanie Seneff and Joseph Polifroni. 2000. Dialogue Management in the MERCURY Flight Reservation System. In *Proc. of ANLP-NAACL, Satallite Workshop, Seattle, WA, 2000.*

Wade Shen, Brian Delaney, Tim Anderson and Ray Slyh. 2007. The MIT-LL/AFRL IWSLT-2007 MT System. In *Proc. of IWSLT, Trento, Italy, 2007*.

Michel Simard, Nicola Ueffing, Pierre Isabelle and Roland Kuhn. 2007. Rule-Based Translation with Statistical Phrase-Based Post-Editing. In *Proc. of the Second Workshop On Statistical Machine Translation, ACL, Prague, Czech Republic, 2007.*

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese Syntactic Reordering for Statistical Machine Translation. In *Proc. of EMNLP, Prague, Czech Republic, 2007*.

Chao Wang and Stephanie Seneff. 2007. A Spoken Translation Game for Second Language Learning. In *Proc. of AIED, Marina del Rey, California, 2007.*

Chao Wang and Stephanie Seneff. 2006. High-quality Speech Translation in the Flight Domain. In *Proc. of Interspeech, Pittsburgh, Pennsylvania, 2006*.

Kenji Yamada and Kevin Knight. 2001. A syntax based statistical translation model. In *Proc. of ACL 2001.*

Dongdong Zhang, Mu Li, Chi-Ho Li and Ming Zhou. 2007. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. In *Proc. of EMNLP, Prague, Czech Republic, 2007.*

Ying Zhang, Stephan Vogel, Alex Waibel. 2004. Interpreting Bleu/NIST scores: How much improvement do we need to have a better system? In *Proc. of LREC, Lisbon, Portugal, 2004.*

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proc. of the Workshop on Syntax and Structure in Statistical Translation, HLT-NAACL, Rochester, NY, 2007.*

Victor Zue, Stephanie Seneff, James Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen and Lee Hetherington. 2000. JUPITER: A Telephone-Based conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing, 8(1): 85-96.*