Combining Linguistic and Statistical Methods for Bi-directional English Chinese Translation in the Flight Domain

Stephanie Seneff, Chao Wang, and John Lee
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA 02139, USA
{seneff,wangc,jsylee}@csail.mit.edu

Abstract

In this paper, we discuss techniques to combine an interlingua translation framework with phrase-based statistical methods, for translation from Chinese into English. Our goal is to achieve high-quality translation, suitable for use in language tutoring applications. We explore these ideas in the context of a flight domain, for which we have a large corpus of English queries, obtained from users interacting with a dialogue system. Our techniques exploit a pre-existing English-to-Chinese translation system to automatically produce a synthetic bilingual corpus. Several experiments were conducted combining linguistic and statistical methods, and manual evaluation was conducted for a set of 460 Chinese sentences. The best performance achieved an "adequate" or better analysis (3 or above rating) on nearly 94% of the 409 *parsable* subset. Using a Rover scheme to combine four systems resulted in an "adequate or better" rating for 88% of *all* the utterances.

1 Credits

This research is supported by ITRI in Taiwan and by the Cambridge MIT Institute. The authors would like to thank Chihyu Chao, Danwen Chen, Mitch Peabody, and Han Shu for their help with data preparation, particularly manual translation of English queries, and for rating translation outputs.

2 Introduction

For over two decades, our group at MIT has been developing multilingual conversational systems as a natural user interface to on-line databases. Within the last few years, our interest has broadened towards the idea of configuring these dialogue systems as a means for a student of a foreign language to practice and perfect conversational skills through spoken interaction with the system. In this language-learning mode, the student would be able to obtain *translation assistance* at any time over the course of a dialogue, by simply speaking a sentence in their native language with equivalent meaning. The system is then tasked with the challenging requirement to provide a fluent translation of their utterance.

Two application domains where we have invested considerable previous effort, both towards multilingual dialogue interaction and towards translation assistance between English and Chinese, are the weather domain (Zue et al., 2000) and the flight domain (Seneff and Polifroni, 2000). We have previously reported on various strategies for achieving high quality and enhancing coverage and robustness for bidirectional speech translation in the weather domain (Wang and Seneff, 2006a; Lee and Seneff, 2005) and for translation from English to Chinese in the flight domain (Wang and Seneff, 2006b). This paper is focused on the specific (new) task of translating from Chinese to English in the flight domain.

We have found in general, as might be expected, that the flight domain is considerably more difficult than the weather domain, due to the much larger number of attributes that can be specified, as well as

the linguistic complexity, in terms of multiple predicates and compound/complex clauses. Fortunately, we have available to us a very valuable resource, which is a set of over 20,000 spoken queries in English that were collected during previous data collection efforts.

In this research, we are interested in addressing the following two questions: (1) How can we exploit the existing English corpus, along with an existing English-to-Chinese translation system, to help in the development of a Chinese-to-English translation system? and (2) How can we effectively combine linguistic and statistical methods to produce a system that exploits the strengths of both approaches?

In the remainder of this paper, we first describe the various strategies that we devised for the Chinese-to-English translation task. We then discuss our evaluation procedure, which is based on manual evaluations of manually provided translations of a set-aside English corpus. After a section detailing our experimental results, we conclude with a discussion of related research and a look to the future.

3 Approach

3.1 Phrase-based SMT

Given a corpus of English sentences within the flight domain and a reasonably high quality English-to-Chinese translation system (Wang and Seneff, 2006b), we can easily generate a parallel English-Chinese corpus, which can be used to train a statistical machine translation (SMT) system. This allows us to quickly develop a reverse Chinese-to-English translation capability, using publicly available SMT tools for training and decoding (Och and Ney, 2003; Koehn, 2004).

As illustrated in Figure 1, we can automatically generate a corpus of English-Chinese pairs from the same interlingual representation by parsing our English corpus and then paraphrasing each utterance into both English and Chinese. It is our belief that the English paraphrase is preferred over the original English sentence because it is likely to be more consistent with the Chinese paraphrase. Furthermore, the English paraphrases usually remove disfluencies present in the original sentences (speech transcriptions), which we do not want the SMT system to capture in its translation models.

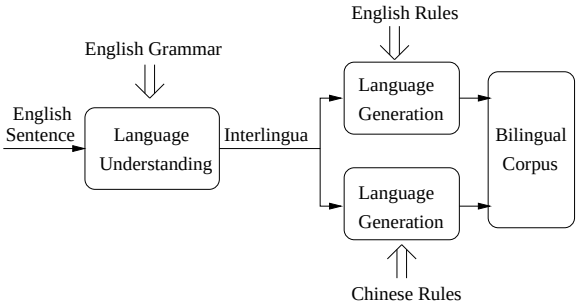

Figure 1: Schematic of technique to automatically generate a synthetic bilingual corpus, for training a statistical translation system.

Once this bilingual corpus is prepared, it can be used to train a statistical machine translation (SMT) system. For this we made use of a state-of-the-art phrase-based MT system developed by Philipp Koehn (Koehn et al., 2003; Koehn, 2004).

3.2 Formal Translation

In parallel, we have developed a linguistic method to translate the generated Chinese corpus back into English. We can use the English grammar rules as a reference to generate comparable Chinese grammar rules, such that there is considerable uniformity in the resulting meaning representations. Ideally, the two languages would produce an identical "interlingua" for sentences with equivalent meaning, and the generation rules would be agnostic to the input language. Our interlingual representation, in contrast to many other interlingual approaches, captures both syntactic and semantic information within a hierarchical structure. Through the device of a trace mechanism (Seneff, 1992) we are able to achieve a strong degree of parallelism in the meaning representations derived from both languages. For example, wh-marked NP's in English are restored to their deep-structure location in the clause, as are temporals and locatives in Chinese.

Figure 2 shows examples of the interlingual meaning representation automatically derived from a parse analysis for an English sentence and a Chinese sentence of equivalent meaning. As seen in the figure, syntactic structure is encoded in the hierarchy, with structural "frames" representing three principal linguistic categories: "{c }" = *clause*, "{q }" = *topic (or noun phrase)*, and "{p }" = *pred-*

meaning representation from the English input:

```
{c wh_question :rhet "there" :auxil "link"
  :topic {q flight :trace "what" :number "pl"
            :pred {p leave  :mode "ing"
                     :pred {p temporal :topic {q weekday :name "monday" }
                              :prep "on"
                              :pred {p at_time  :prep "after"
                                       :topic {q clock_time
                                                 :pred {p clock_hour :topic 5 }
} } } } } }
```

meaning representation from the Chinese input:

```
{c wh_question :rhet "there"
  :topic {q flight  :trace "what"
            :pred {p leave
                     :pred {p temporal :topic {q weekday :name "monday" }
                              :pred {p at_time :prep "after"
                                       :topic {q clock_time
                                                 :pred {p clock_hour :topic 5 }
} } } } } }
```

Figure 2: Comparisons of meaning representations for a pair of equivalent English and Chinese sentences: "what flights are there leaving after five o'clock on monday" and "you3 shen2 me5 xing1 qi1 yi1 wu3 dian3 zhong1 yi3 hou4 chu1 fa1 de5 hang2 ban1" (literally "have what monday five o'clock after leave <PARTICLE> flight.") Features missing from the Chinese analysis are highlighted in bold fonts.

icate. The trace mechanism allows the forward-moved constituent, "what flights," in the English sentence to be represented identically to the unmoved "shen2 me5 ... hang2 ban1" in Chinese.

Despite the similarity between the interlingual frames generated from the English sentence and its Chinese counterpart, it is often the case that many English syntactic features are impoverished in the Chinese representation. Generally, Chinese does not express explicitly the definiteness ("a" vs. "the") and number (singular/plural) of nouns, or the mode/number of verbs ("leave" vs "leaves" vs "leaving"). In the example shown in Figure 2, several linguistic features are missing from the Chinese representation, such as the `:mode "ing"`, `:number "pl"`, and `:prep "on"`. Hence, when the parse analysis is correct, the linguistic structure of the resulting generation string is typically well-structured, except that it may still be apparent that it is a "non-native" rendering of the sentence. Figure 3 illustrates some examples of English translations derived from Chinese input, when the missing features are not adequately handled by generation rules.

Our generation system includes a preprocessor stage (Cowan, 2004) which can augment an interlingua frame with syntactic and semantic features specific to the target language, for example, deciding between definite/indefinite articles for noun phrases. However, predicting definiteness in English is not straightforward, and therefore prone to error. In this paper, we have sought new methods to address these hard problems, which will be described in more detail in Section 3.3.

A further problem is the classical "PP-attachment" problem, or parse ambiguity. The grammars we have developed for both Chinese and English, although they utilize a powerful probability model (Seneff, 1992), are still capable of producing erroneous parse analyses, which can lead to incorrect translations. We have intentionally developed grammars based on syntactic structure, so that they can easily be ported to other domains. However, the flight domain is far more challenging than the weather domain in terms of ambiguous parses, and there is no guarantee that the correct parse analysis will achieve the highest score.

After exploring a number of options, we found that it was productive to empower the developer to manually provide a set of domain-dependent "frame rewrite rules," to specify a reorganization of the interlingua, when necessary, to reflect a semantically more plausible alternative. While these rules are domain specific, their syntax is quite straightforward,

original English	"impoverished" English
a later one	later one
the earlier flight	earlier flight
departing in the morning	depart in the morning
i would like to leave on friday	i want to leave friday
i am looking for flights from memphis to london	i find flight from memphis to london

Figure 3: Examples of English sentences and their paraphrases after an English-to-Chinese-to-English translation cycle.

```
{c rule   :in ( give, list, show ... )
          :contents ( temporal, fare_class, ... )
          :to flight }
```

Figure 4: Example of a structure rewrite rule to correct misplaced attachment of a "temporal" or a "fare_class" predicate.

and they allow the developer to gain control over the problem of erroneous parses. Figure 4 shows an example of such a rule, which specifies a list of predicates which, if attached to the verb in the original frame, should be moved to a flight noun phrase, if it exists, essentially correcting an inappropriate PP-attachment in the parse tree. Currently, we make use of about 30 such rewrite rules.

For language generation, we use a generation system, GENESIS (Baptist and Seneff, 2000), that operates from rule-templates to generate surface strings from the interlingual meaning representation. Since the English and Chinese grammars share a common convention in their rules, and because of the trace mechanism to regularize the structure, we were able to use exactly the same set of language generation rules for Chinese-to-English generation as were used for English-to-English generation.

GENESIS uses 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 the constituents in the generated string. Generation begins by looking up the rule-template for the top-level clause constituent of the frame, and proceeds recursively through the rules, concatenating substrings contributed by the frame's constituents to form each frame's generation string.

Typically, there is a default rule for each of the three main constituent categories: clause, noun phrase, and predicate, where "predicate" refers broadly to prepositional phrases and adjective phrases as well as verb phrases. Any constituent can privatize its own unique generation rule, if needed, and similar noun or verb phrases can share the same generation rule by forming groups (e.g., "show" and "tell"). Variability in the surface form can be achieved by randomly selecting among alternative rules and lexical entries, although we have not yet exploited this capability in the flight domain generation. A recent research effort is to improve the portability of the rules. For example, a set of pre-specified generation rules for all temporal expressions can just be included for any domain that uses temporal expressions. The current rule set for the flight domain in English contains about 140 unique rules.

As mentioned earlier, we have developed an independent preprocessor stage for GENESIS (Cowan, 2004) that can augment an interlingual frame with features appropriate for a given target language. The syntax of its rules is very similar to that of the main processor, but its effect is to add syntactic and semantic features to the frame, thus simplifying the burden placed on the main processor.

3.3 Post-editing Methods

As noted above, Chinese utterances are typically impoverished in ways that could lead to classic mistakes that are common for a native Chinese speaker. Since our system would be providing examples of "correct" English usage to Chinese speakers, it is important for it to provide accurate accounting for these difficult-to-master aspects of the language. Thus we sought a methodology to correct such deficiencies, which would almost certainly require a statistical approach.

There are at least two logical candidates for this: (1) use an "English-to-English" SMT system to map

"bad" English to "good" English, and (2) use an n-gram language model, along with a parsing step to select preferentially for long-distance constraints (Lee and Seneff, 2006). Statistical translation systems have the distinct advantage that they require very little manual effort. The main requirement is that they be provided with a high quality parallel corpus. We had in hand a very easily obtainable corpus: use formal methods to translate from English to Chinese to English, and to translate from English directly to English. Figure 5 shows the generation of a parallel corpus used for training an SMT system for correction: the English-to-English translation is the "output language," whereas the more errorful English-to-Chinese-to-English translation is the "input language."

A second method we decided to explore is one that is a natural extension of our prior research on correcting ill-formed sentences produced by students of a second language (Lee and Seneff, 2006). We again exploit the direct English-to-English translation system to provide a model of "correct" usage. But instead of a complete SMT system, we use simple rewrite rules to license only a small set of carefully selected alternative forms. For instance, if there is expected to be ambiguity in the choice of the article, the generation system can produce "(a || the || NULL)" as the output, and allow a later stage to make the decision statistically. The schema we developed thus uses an overgenerate-and-select paradigm to determine appropriate usage of function words and verb inflectional endings.

This process is illustrated in Figure 6. The formal language generation system produces alternative choices for a subset of its string outputs. The resulting word graph is processed through a standard class n-gram language model, trained on the outputs of the English-to-English "translation" system. The resulting N-best list (N = 10) of candidates are then parsed with the natural language grammar for English. The hypothesis that yields the highest combined score (n-gram and parse) is selected.

4 Example Outputs

In order to illustrate how we exploited our English-only flight corpus to develop and statistically train our translation systems, we will walk through a series of derived paraphrase variants produced from a single example taken from the corpus. Figure 7 shows outputs from various stages of processing for the English sentence, "I need a one way flight from boston to pittsburgh on april twenty eighth in the morning." The English-to-English paraphrase (2) is slightly modified, with the ordering of the date and time reversed, and the cardinal date converted to an ordinal date. A separate "tagged English" paraphrase (3) is created to train the statistical class n-gram language model. It has all numbers and dates converted to the appropriate printed form, and associated with a tag ("<day>" in the figure) that provides class names for the class n-gram. This technique allows the language model to better generalize from the training set. The English-to-Chinese paraphrase (4) provides training utterances for both the SMT system and the Chinese parsing grammar. The Chinese string is converted back to English using the same English language generation rules that are used for English-to-English paraphrases (5). Finally, the Chinese-to-multi-English paraphrase (6) has the same format as the tagged English, but, in addition, has alternative choices for a select set of function words, such as articles and prepositions. It is used as input to the n-gram for the post-editing correction phase.

While it may seem that maintaining all of these different variants of the English paraphrases is cumbersome, it turns out that they all share a common lexicon and set of language generation rules, as illustrated in Figure 8. All of the variants are produced through a final string edit on the original output string, where each of the distinct English forms (tagged-English, multi-English, and standard English) has its own unique string rewrite rules. For example, the original generation paraphrase might contain a "$*$quant$*$" indicating a location where a quantifier might be present, but it is wary of committing to a particular one. This is deleted by a string rewrite rule for both standard English output and tagged English output, but it is rewritten as (a || an || the || NULL) for the multi-English output, allowing the statistical post-processing to select the most probable one. The tags for the n-gram language model are provided in the original generation output, but automatically deleted by the standard English rewrite rules.

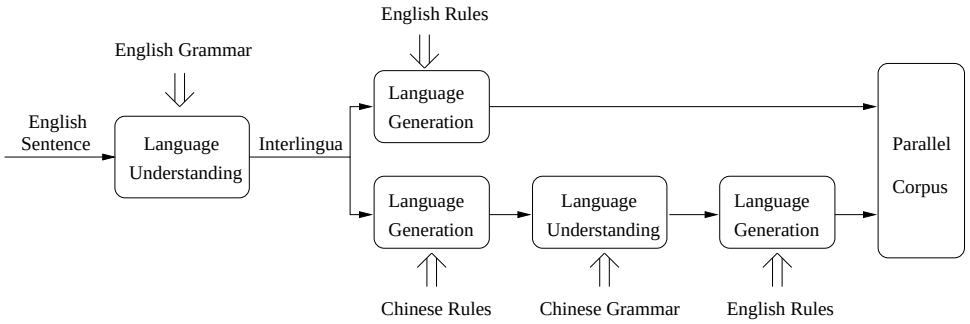

Figure 5: Schematic of technique to automatically generate a parallel corpus to train a statistical translation system for correcting translations generated by the formal system.

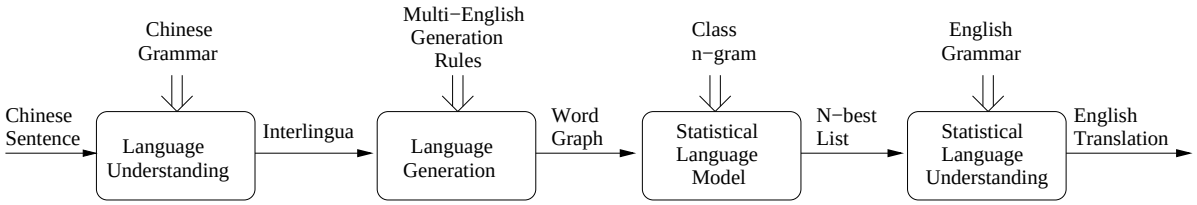

Figure 6: Diagram of translation procedure in which a formal parse-generate method overgenerates translation candidates. The highest scoring parsable candidate from an N-best list obtained through a class n-gram language model is selected as the final translation string.

(1)	Original English	I need a one way flight from Boston to Pittsburgh on April twenty eighth in the morning
(2)	English-to-English Paraphrase	I need a one way flight from Boston to Pittsburgh in the morning on April twenty eight
(3)	Tagged English Paraphrase	I need a one way flight from Boston to Pittsburgh in the morning on April 28/<day>
(4)	English-to-Chinese Paraphrase *Literal Translation (for reference)*	wo3 xu1_yao4 si4_yue4 er4_shi2 ba1 hao4 shang4_wu3 cong2 bo1_shi4_dun4 dao4 pi3_zi1_bao3 dan1_cheng2 de5 hang2_ ban1 *I need April twenty two afternoon from Boston to Pittsburgh* <PARTICLE> *one way flight*
(5)	Chinese-to-English Translation	I need one way flight from boston to pittsburgh in the morning April twenty eight
(6)	Chinese-to-Multi-English (Tagged)	I need (a || an || the || NULL) one way (flight || flights) from Boston to Pittsburgh in the morning (on || NULL) April 28/<day>

Figure 7: Examples of various paraphrases produced from a single English utterance in our data collection corpus, for use by various configurations of the Chinese-to-English translation system.

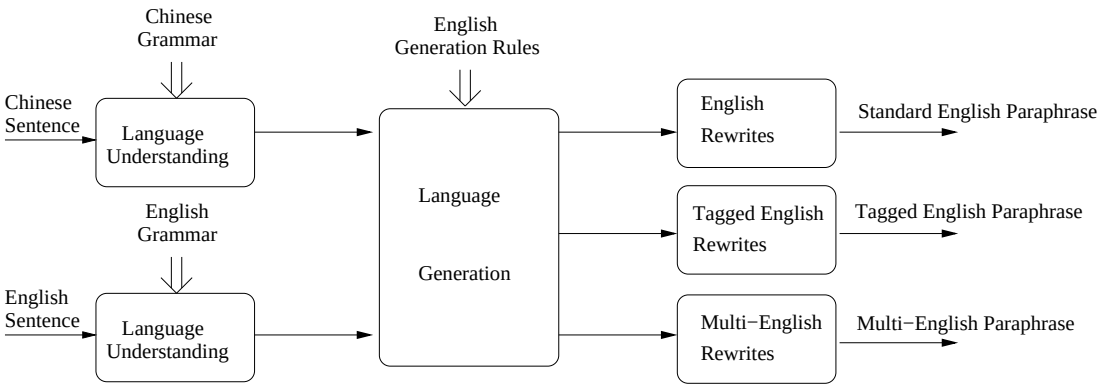

Figure 8: Illustration of System configurations for generating all the different forms of English and marked-up English outputs used by the various systems. Note that all configurations make use of the same set of English generation rules.

5 Evaluation Procedure

In this section, we describe in detail the set of system configurations we explored, and the procedures that were used to compare and contrast the performance of the various competing strategies. As illustrated in Figure 9, we begin with two basic systems: (1) a "standard" SMT system (system I), and (2) a "standard" interlingua-based translation system (System II). We then applied two different methods to attempt to further improve the outputs of the interlingual system: (3) process the outputs through an SMT system trained on aligned "bad" English/"good" English pairs (System III), and (4) process an overgenerated set of interlingual outputs through a selection process involving n-gram statistics followed by selection by parsing using the standard English grammar (System IV). Systems II, III, and IV all depend on the Chinese understanding system to produce an interlingua for an input sentence, and hence would result in a null translation output upon parse failure. Thus we will also report a performance achieved by combining the best of Systems II-IV with System I as a backup mechanism.

Since we did not have unseen Chinese sentences in the flight domain for evaluation, we created the test data by asking bilingual speakers to manually translate English flight queries into Chinese. Using our English-to-Chinese translation capability (Wang and Seneff, 2006b), we automatically generated a corpus of English/Chinese sentence pairs from our available 20,000 English flight queries, To ensure separation of training and test data, we obtained about 10,700 *unique* parsable English queries from the original corpus, and divided that into three sets: training (5/6 of total data), development (1/12) and test (1/12). The three sets have similar distributions in sentence length, ranging from 1 to around 30 words, with an average of 7.3 words per sentence. The training set was used to generate various parallel corpora for training the statistical components in our system, as well as for discovering coverage gaps in the formal parsing and generation rules. Both the development and test sets were translated into Chinese manually to provide data for system refinement and evaluation.

Translations for the test utterances are generated under each system configuration, and grouped by the original Chinese input. Due to the effort required to conduct manual ratings, and also so that we could set aside data for future evaluation experiments, we used about half of the reserved test data in our evaluation. We generated 1082 unique translations for the 460 input sentences. A bilingual judge rated all the translations for each sentence collectively, without explicit knowledge of the origin of each translation. These ratings were checked by a second native English speaker. A scale of 1 to 5 is used in the ratings, with 5 being perfect, 3 being acceptable (correct syntax and semantics, but inappropriate usage of function words or slightly odd constructs), and 1 being incorrect (either semantic content or syntactic structure). Ratings 2 and 4 are used sporadically for those cases that fall between the major categories.

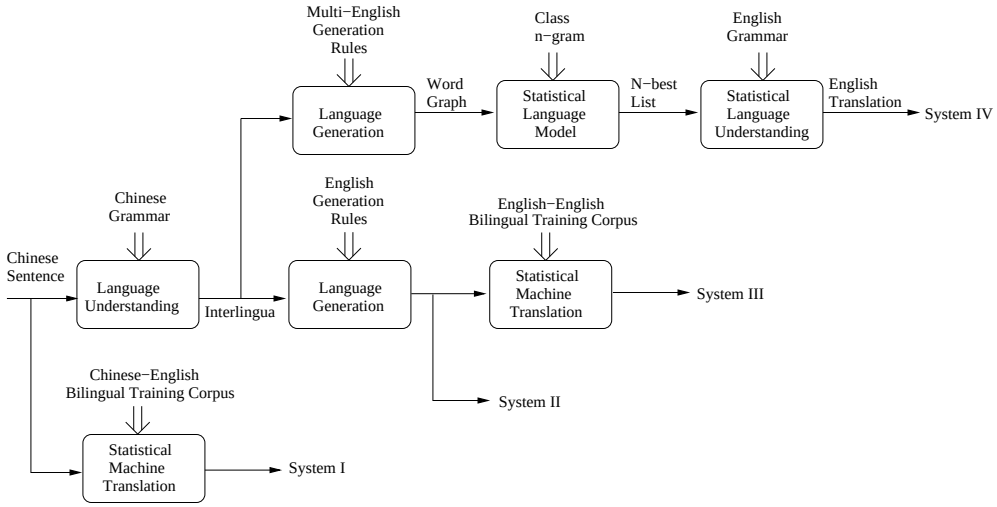

Figure 9: Four system configurations that were used for the evaluation experiments.

	5	4	3	2	1	F
I	277	6	45	12	120	0
II	270	28	83	8	20	51
III	333	10	33	7	26	51
IV	322	15	47	8	17	51
III+I	345	11	40	13	51	0
IV+I	334	16	54	14	42	0
Rover	358	12	36	10	44	0

Table 1: Human ratings of translation outputs of various system configurations on a set of 460 Chinese inputs.

	5	4	3	2	1	Average
I	265	5	38	6	95	3.83
II	270	28	83	8	20	4.27
III	333	10	33	7	26	4.51
IV	322	15	47	8	17	4.51
Rover	346	11	29	4	19	4.62

Table 2: Human ratings of translation outputs of various systems on the subset of 409 parsable Chinese inputs.

Figure 10 shows some example translation outputs and the corresponding human ratings.

6 Results and Discussions

Table 1 summarizes the ratings each system received on the 460 inputs in the test data, while Table 2 analyzes the common subset of 409 sentences for which all systems generated a non-empty translation.

The statistical system (I) and formal system (II) received similar numbers of "perfect" ratings, although the number of "bad" outputs (those receiving a rating of 1 or 2) is significantly higher in the statistical system. It is our observation that the SMT system tends to produce spurious words in its translation outputs when it falls apart on novel inputs.

The two post-correction methods both improved the formal translation outputs. It is interesting to observe that the statistical system is able to increase the number of "perfect" translations from 270 to 333, a 23.3% improvement. A comparison of the translations before and after the correction shows that the SMT correction system is able to learn systematic differences between the English paraphrases from Chinese inputs vs. from English inputs. However, because the correction is purely data-driven, it suffers from the same problem the baseline SMT system encountered: the number of "bad" translations also increased after the correction. On the other hand, the post-correction method in System IV adopted a much more conservative strategy: it constrains the possible types of corrections via formal generation methods (by generating a word graph), while using statistics to choose the preferred solution within that restricted space. It was rarely the case that the correction resulted in a worse translation output. In fact, it is able to achieve "adequate"

1.	gao4 su4 wo3 wu3 dian3 ban4 yi3 hou4 de5 hang2 ban1 .	
	(*Literally:* tell me five o'clock half after <PARTICLE> flight.)	
1a.	tell me the flight at five after	1
1b.	tell me flight after five thirty	3
1c.	tell me about flights after five thirty	5
1d.	tell me the flight after five thirty	4
2.	wo3 yao4 xia4 yi1 ban1 cong2 kang1 zhou1 ha1 te4 fu2 dao4 ya4 te4 lan2 da4 de5 ban1 ji1 .	
	(*Literally:* i want next <PARTICLE> from connecticut hartford to atlanta <PARTICLE> flight.)	
2a.	i want to go from hartford connecticut to atlanta the next one	1
2b.	i want a next flight from hartford connecticut to atlanta	3
2c.	i would like the next flight from hartford connecticut to atlanta	5
2d.	i want the next flight from hartford connecticut to atlanta	5
3.	ni3 ke3 yi3 gao4 su4 wo3 zao3 yi1 dian3 de5 hang2 ban1 ma5 ?	
	(*Literally:* you can tell me earlier <PARTICLE> flight <Q PARTICLE>?)	
3a.	could you tell me the earlier flight that	1
3b.	can you tell me earlier flight	3
3c.	can you tell me about the earlier flight	5
3d.	can you tell me the earlier flight	5

Figure 10: Translation outputs and human ratings for three Chinese inputs. (5=Perfect, 3=Acceptable, 1=Incorrect)

or better translation on nearly 94% of the parsable subset. We only experimented with a very limited set of corrections (i.e., articles, noun singular/plural forms, verb modes, and prepositions before date and time expressions). It is conceivable that the space can be expanded to cover other systematic forms of awkwardness in the translation. Both correction systems (III and IV) achieved the same average ranking score.

We were able to use a simple "Rover" scheme to select among the four candidate outputs, based on maximizing parse score, to achieve a performance that was better than that of any single system, as shown in the last row in Table 1 and Table 2.

7 Related Research

(Langkilde and Knight, 1998) were pioneers in introducing the idea of using a linguistic generation system to overgenerate a set of candidate hypotheses for later selection by a statistical evaluation. Their *Nitrogen* language generation system processes from an underspecified semantic input, which they call an "AMR" (abstract meaning represention). It outputs a word graph representing a very large list of alternative realizations of the syntactic sugar missing from the AMR. An n-gram language model then selects the most plausible realization. As in our work, they generate both singular and plural forms of underspecified nouns, and allow the statistical system to select the more likely form, given the preceding context. But they also allow more global rearrangements of the structure, including active versus passive voice realization of the verb, for example. Their generation formalism resembles ours, in that a set of "keyword rules" map semantic and syntactic roles to grammatical constructs. However, because their representation is missing explicit syntactic information, the hypothesized syntactic organization is encapsulated in the (alternative) rules rather than directly in the AMR. Their notion of an "AMR recasting rule" is similar to our corrective rule rewrite mechanism, and allows an original AMR to be cast into related AMR's.

A number of projects have involved language translation in the flight domain, particularly involving spoken language translation (Gao et al., 2002; Ratnaparkhi, 2000; Rayner and Carter, 1997), probably due to the existence of large English speech corpora obtained from spoken dialogue systems (such as ATIS and the Darpa funded Commmunicator Program). (Rayner and Carter, 1997) advocate combining rule-based and statistical methods to achieve robust and efficient performance within a linguistically motivated framework. Their system translates from English into Swedish and French, and uses statistical methods for word-sense disambiguation. They argue for utilizing expertise to develop generic grammar rules covering the core linguistic constructions

of a language, which is similar to our approach to grammar development. Whenever the formal parsing method fails to deliver a solution, a less sophisticated back-off based on direct mapping from words and phrases is utilized.

(Gao et al., 2002) have developed a reversible spoken language English/Chinese capability, which is based on a strictly semantic representation of the input sentence, serving as an interlingua. This language-independent tree-structure representation is derived via a statistical decision-tree model obtained through automatic training from an annotated corpus of spoken queries. A Maximum Entropy based language generation approach, trained from the same annotated corpus, is adopted for generating the target language output, providing both attribute ordering and lexical choice (Ratnaparkhi, 2000). Ratnaparkhi concedes that trainable approaches, while avoiding the expense of handcrafted rule development, are less accurate than rule-based approaches.

8 Future Work

In examining the bad-english to good-english SMT translation outputs, it has become apparent to us that a viable approach for improving the linguistic system is to use the SMT system's proposed edits as a mechanism for identifying problematic linguistic translations. Corrections can be made either directly in the formal rules or in the statistical post-processing step, on a case-by-case basis. Another direction we plan to explore is developing a strategy for deciding whether or not to trust a proposed translation, which will allow the system to apologize rather than risking presenting the user with false information about the linguistic constructs of the language being learned. Such a decision could be based on the quality of a parse analysis of the translation, or on an n-gram score, for example. In future work, we plan to extend the methodologies developed for the flight domain into other domains more appropriate for language learning.

References

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

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

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

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

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. AMTA*, Washington DC.

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *P roc. COLING-ACL*.

J. Lee and S. Seneff. 2005. Interlingua based translation for language learning systems. In *Proc. ASRU*.

J. Lee and S. Seneff. 2006. Automatic grammar correction for second language learners. To appear in *Proc. Interspeech*.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

A. Ratnaparkhi. 2000. Trainable methods for surface natural language generation. In *Proc. HLT-NAACL*.

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

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

S. Seneff. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1):711–714.

C. Wang and S. Seneff. 2006a. High-quality speech-to-speech translation for computer-aided language learning. To appear in *ACM Transactions on Speech and Language Processing*.

C. Wang and S. Seneff. 2006b. High-quality speech translation in the flight domain. To appear in *Proc. Interspeech*.

V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington. 2000. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96.