High-Quality Speech-to-Speech Translation for Computer-Aided Language Learning

CHAO WANG and STEPHANIE SENEFF
MIT Computer Science and Artificial Intelligence Laboratory

This article describes our research on spoken language translation aimed toward the application of computer aids for second language acquisition. The translation framework is incorporated into a multilingual dialogue system in which a student is able to engage in natural spoken interaction with the system in the foreign language, while speaking a query in their native tongue at any time to obtain a spoken translation for language assistance. Thus the quality of the translation must be extremely high, but the domain is restricted. Experiments were conducted in the weather information domain with the scenario of a native English speaker learning Mandarin Chinese. We were able to utilize a large corpus of English weather-domain queries to explore and compare a variety of translation strategies: formal, example-based, and statistical. Translation quality was manually evaluated on a test set of 695 spontaneous utterances. The best speech translation performance (89.9% correct, 6.1% incorrect, and 4.0% rejected), is achieved by a system which combines the formal and example-based methods, using parsability by a domain-specific Chinese grammar as a rejection criterion.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*; *language generation*; *machine translation*; K.3.1 [**Computers and Education**]: Computer Uses in Education—*Computer-assisted instruction*

General Terms: Algorithms, Experimentation, Languages

Additional Key Words and Phrases: Speech translation, machine translation, dialogue systems, computer-aided language learning

1. INTRODUCTION

Speech-to-speech translation is a challenging problem due to the nonstandard/informal style typically associated with spontaneous speech as well as errors caused by automatic speech recognition. To achieve a usable performance level, most speech translation systems reported in the literature operate within more or less restricted domains [Alshawi et al. 2000; Levin et al. 2000; Ney et al. 2000; Gao et al. 2002; Casacuberta et al. 2004].

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

Author's address: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139; email: wangc@csail.mit.edu.

Knowledge-lean statistical machine translation approaches [Brown et al. 1993] are nearly universally embraced for the task of unrestricted text translation, perhaps because it is more difficult to effectively exploit knowledge in the broad domain. In restricted domains, however, it becomes possible to explore the use of deep linguistic and even world knowledge in the translation process. For instance, Levin et al. [2000] uses a semantic interlingua called *Interchange Format*, and both the parsing and generation components use formal grammar rules. Gao et al. [2002] also uses semantic information as an interlingua; however, a statistical framework is adopted both for parsing and generation. Some speech translation systems pursue a purely statistical machine translation approach [Alshawi et al. 2000; Ney et al. 2000; Bangalore and Riccardi 2002; Kuman and Byrne 2003; Casacuberta et al. 2004], while emphasizing a tighter interface between the recognition and the translation components that is typically achieved by way of lattice-based decoding algorithms or finite-state models.

Our goal is to develop a speech-to-speech translation framework to be used as a language tutor in computer-aided language learning (CALL) systems [Seneff et al. 2004]. This presents special challenges because the quality of the translation must be essentially perfect to avoid teaching the student inappropriate language patterns. To achieve this goal, we have proposed an interlingua-based translation framework in which formal rule-based translation is augmented with an example-based translation mechanism for improved robustness [Wang and Seneff 2004]. The formal approach consists of separate parsing and generation steps, mediated by a *semantic frame*, our instantiation of the interlingua, which is a hierarchical structure encoding detailed semantic and syntactic information of the input sentence. The formal method is able to produce very high-quality translation for inputs within the coverage of the rules. However, it has the disadvantage of hard failure on novel inputs, that is, when a user uses expressions unforeseen by the rule developer or when the inputs are ill-formed due to nonstandard usage or automatic speech recognition (ASR) errors. To overcome this deficiency, we devised a more robust translation-by-example mechanism in which the semantic frame is reduced to a list of key-value (KV) pairs representing critical semantic information of the input sentence. The KV information is then used to match against a KV-indexed translation corpus to find an appropriate candidate. Since KV information can usually be extracted even with keyword spotting, this method is much more robust to new linguistic patterns and ill-formed sentences. The KV-indexed translation corpus can be automatically constructed by running the formal translation procedure on a set of sentences as described later in this article.

With the formal rule-based translation system, we are able to automatically create a large bilingual parallel corpus from a monolingual (English) one, collected over the years from a publicly available dialogue system in the weather domain [Zue et al. 2000]. This has enabled us to explore other translation frameworks such as purely statistical methods. Thus a second goal of this article is to benchmark our system against other approaches to machine translation, in particular statistical methods. Statistical machine translation systems have achieved state-of-the-art performance for general-domain translation. We are interested in finding out whether they can achieve competitive performance for narrow-domain applications given adequate in-domain data for training.

In the remainder of the article, we first give an overview of our translation framework in Section 2. We then describe each technology component in more detail. Section 3 covers the interlingua representation which is derived by parsing the input sentence. Sections 4 describes the natural language generation component. Section 5 describes the example-based translation framework which includes a class-based back-off mechanism. Empirical results on manually and automatically derived speech transcriptions are reported in Section 6 with comparison to a baseline system obtained by training a phrase-based statistical machine translation model for our application domain. Related research is discussed in Section 7, followed by conclusions and future work in Section 8.

2. SYSTEM OVERVIEW

Our translation framework adopts the interlingua approach and is integrated with the dialogue system development by means of a shared meaning representation which we call a *semantic frame*. Given an input sentence, a parse tree is derived, and critical syntactic relations and semantic elements in the parse tree are extracted. The resulting semantic frame can be used to generate key-value information for querying the dialogue system and to generate a sentence in the original language (paraphrasing) or in a different language (translation). We adopt a formal approach in both the parsing and generation components, while emphasizing portability of the grammar and generation rules to new domains [Rayner and Carter 1997]. The parser [Seneff 1992b] utilizes a context-free grammar augmented by a feature unification mechanism. It also automatically acquires a probability model for the context-free rules by parsing a corpus of unlabeled data. The generation is controlled by a set of rules and a context-sensitive lexicon which can be fine-tuned to achieve high quality.

Our dialogue-based tutoring system employs two grammars, one to parse the native language (L1) for translation, and one to parse the foreign language (L2) for dialogue processing. We can make use of the L2 grammar to achieve some quality assurance on the translation outputs. If the generated translation fails to parse under the L2 grammar, we resort to an example-based method in which semantic information encoded as key-value pairs is used to search a precompiled indexed L2 corpus for a suitable candidate. If both methods fail, the system will prompt the student to rephrase. We think that a null output is better than an erroneous one, given the intended use of our system. The example-based mechanism complements the rule-based generation in that it tends to be more robust for ill-formed inputs [Levin et al. 2000; Frederking et al. 2002]. Figure 1 illustrates the flowchart of the translation procedure.

Figure 2 summarizes the formal rule-based translation system, configured for the scenario of a native English speaker learning Chinese. Translation is achieved by parsing the English sentence to obtain a semantic frame and generating from the semantic frame a Chinese surface string. As illustrated in the

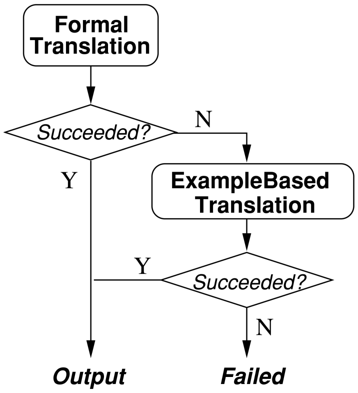

Fig. 1. Flowchart of the translation procedure.

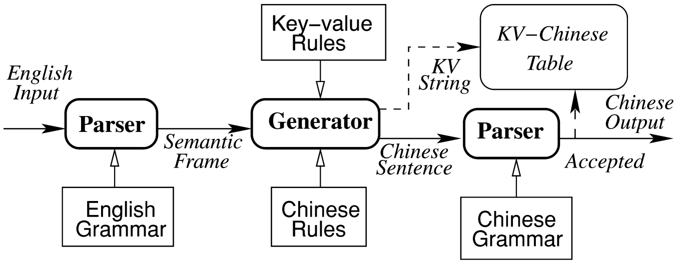

Fig. 2. Schematic diagram for formal translation method. (KV = Key Value)

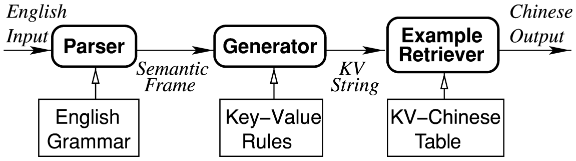

Fig. 3. Schematic diagram for example-based translation method. (KV = Key Value)

figure, this procedure can also be used to automatically produce an indexed Chinese corpus from a collection of English sentences to serve as a translation memory for the example-based method. In this case, a list of key-value pairs is automatically generated from the semantic frame for each English input which is then used to index the corresponding Chinese surface string in the translation table. The translation table can also be augmented with any available original Chinese data in which case the key-value index can be derived using the Chinese grammar for parsing into a common semantic frame representation.

The example-based translation method is very straightforward, given the availability of a KV-indexed translation corpus. Figure 3 summarizes the procedure, again for the scenario of translating English sentences into Chinese. The English sentence is parsed to obtain a semantic frame from which a KV string is generated using trivial generation rules. The KV string is used to retrieve an appropriate Chinese sentence from the translation table as the output.

3. PARSING AND THE INTERLINGUA

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

Constructing a grammar for each conversational domain (lesson plan) is a time-consuming process that requires either large amounts of labelled in-domain data to automatically induce and train a grammar, or linguistic expertise to compensate for the lack of data. In either case, manual effort is unavoidable, either to annotate the data, or to write grammar rules. It is thus appealing to develop a syntax-based grammar that covers a broad domain so that the manual effort would be a one-time investment that can benefit a wide range of applications. The Penn Treebank project [Marcus et al. 1993] exemplifies such a philosophy in which linguistic information (syntactic structure and part-of-speech tags) of naturally-occurring text (Wall Street Journal, the Brown Corpus, etc.) is annotated. Facilitated by the Treebank corpus, a number of data-driven statistical parsers have been developed, achieving high accuracy on the same type of data [Collins 1997]. Unfortunately, the performance typically does not carry over to conversational domains where both the style and the semantic content are quite different from those of the Wall Street Journal. There is no similar effort like Penn Treebank in collecting and annotating conversational speech except for some limited effort in the SwitchBoard domain [Godfrey et al. 1992] and the ATIS air travel information domain [Price 1990]. In human-computer dialogue system development, a practical compromise is to train a shallow semantic parser from coarsely (and possibly automatically) labeled data [Tang et al. 2002; He and Young 2003] which usually results in a relatively flat semantic structure. Such a representation, while acceptable for dialogue processing, is generally not adequate for deriving an accurate translation of the input.

Our solution is to construct a generic core grammar that is linguistically rich and easily adaptable for different domains within the dialogue interaction scenario. The grammar is induced semiautomatically, making heavy use of existing resources such as several grammars developed for various dialogue applications [Zue et al. 2000; Seneff and Polifroni 2000] as well as speech corpora collected using those systems. We began by substituting domain-specific nouns, verbs, and adjectives in the collected user data in various domains with their corresponding part-of-speech (POS) tags. We then developed a largely syntax-based core grammar to parse these pseudo sentences. The core grammar

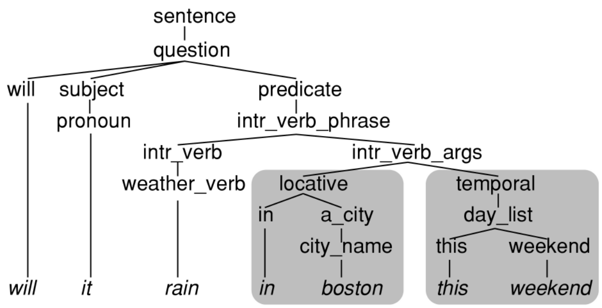

Fig. 4. Parse tree of an example sentence “Will it rain in Boston this weekend?”

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

Fig. 5. Semantic frame of an example sentence “Will it rain in Boston this weekend?”

contains top-level rules specifying detailed syntactic structure with part-of-speech nodes to be expanded into words or word classes specific to a domain. Rules for general semantic concepts such as dates and times are organized into subgrammars that are easily embedded into any domain. Phrases for subgrammars are also extracted from existing speech transcripts which can be re-used to train a statistical language model for the speech recognizer in the new domain [Seneff et al. 2003].

Figure 4 illustrates the parse tree obtained using our core English grammar adapted for the weather domain. The subgrammars to parse the location phrase “in Boston” and the date phrase “this weekend” (highlighted in rectangular shades in the figure) are directly provided by the core grammar. Only domain-specific nouns and verbs (e.g., rain as a weather_verb) need to be entered by a developer.

The parse tree is then mapped to a semantic frame, our instantiation of the interlingua. The term semantic frame came historically from our dialogue system development research; in our current framework, it captures both syntactic and semantic information to enable the generation component to produce a highly accurate translation of the original input. The mapping from parse tree to semantic frame is very straightforward: selected parse nodes are mapped to constituents in the semantic frame (clauses, topics, and predicates, etc.), with their hierarchical relationships in the parse tree preserved in the semantic frame representation. Figure 5 illustrates the semantic frame derived from the parse tree shown in Figure 4.

It is desirable, if not mandatory, to maintain high consistency in the semantic frame derived from different languages to make it a true interlingua. This is challenging when the ordering of major syntactic constituents in different languages differs significantly as exemplified by wh-questions in English and Chinese. Wh-questions in English are formed by preposing the wh-marked constituent to the beginning of the sentence: "*How much snow* will they get?" By contrast, in Chinese, wh-marked noun phrases occur in the *same unmoved* position as their non-wh-marked counterparts:

ta1 men2	hui4	de2	*duo1 shao3*	*yu3*
they	*will*	*get*	*how much*	*snow*

However, Chinese does utilize a similar forward-movement strategy to topicalize locatives and temporals for both statements and questions where the presence of the question particle "ma5" encodes the interrogative form:

ming2 tian1	*bo1 shi4 dun4*	hui4	xia4 yu3	[ma5]
tomorrow	*Boston*	will	rain	[`question-particle`]

This sentence corresponds to a statement "It will rain in *Boston tomorrow*." or a question "Will it rain in *Boston tomorrow*?" depending on the absence or presence of "ma5."

To achieve a homogeneous representation, we utilize a trace mechanism in our parsing algorithm, supporting constituent movement and restoring the moved constituent to its underlying natural position in the derived semantic frame. A complementary mechanism exists in the generation system to manage the reverse process as described in the next section. The parsing rules guiding the movements are included in the core generic grammar and are readily available to all domain-specific grammars without extra effort from the developer.

Any remaining language-dependent aspects in the semantic frame such as features that are present only in certain languages (e.g., articles in English, particles in Chinese, etc.), and special/idiomatic expressions are handled by the generation component as described in the next section.

4. NATURAL LANGUAGE GENERATION

To generate well-formed strings in L2, we utilize the GENESIS language generation framework [Baptist and Seneff 2000]. It works from a lexicon which provides context-dependent word-sense surface strings for each vocabulary item, along with a set of recursive rules that specify the ordering of constituents in the generated string. Variability in the surface form can be achieved by randomly selecting among alternative rules and lexical entries. For example, there are two ways to construct yes-no questions in Chinese. We illustrated the `statement + question-particle` construct earlier, but an equally correct alternative is the so-called `A-not-A` construct:

ming2 tian1	bo1 shi4 dun4	hui4 bu2 hui4	xia4 yu3
tomorrow	*Boston*	*will-not-will*	*rain*

The English example in Figure 4 can be alternatively realized as the Chinese `statement + question-particle` construct or the `A-not-A` construct with additional permutations on the ordering of the date and location phrases. Thus, the same English input processed through the system four times could yield four different variants. This is useful not only to the language student, but also to system development since we can generate a rich set of Chinese sentences for training the language models of the speech recognizer.

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

One of the features of our generation mechanism is the ability to support sharing of generation rules among a large number of elements that follow a common generation pattern. Typically, there is a default rule for each of the three major constituent types: clause, topic, and predicate, that covers the most common generation pattern for each type. Verbs that can take a clause complement are grouped together into a common class, associated with a specialized generation rule that handles the clause complement. For example, if a new verb is introduced into the domain, instantiating generation coverage for itself and its arguments/complements amounts to figuring out which group it logically belongs to and adding its surface form realization to the lexicon. Thus, we feel that the generation rules, while domain-dependent, provide a substantial domain-independent core rule base that would support an efficient port to a new domain, a strategy that is analogous to our approach in parsing.

A difficulty in translating utterances in a database query domain is that most of the utterances are questions which are encoded very differently in different languages. We have given some examples of wh-questions and yes-no questions in the previous section. An even more complicated situation arises in spoken language by a prevalence of polite forms with embedded clauses, "Do you know if it will snow tomorrow in Boston?" In English, an inversion of the main clause is sufficient to encode the interrogative form, whereas in Chinese, when the `A-not-A` construct is used, both the main clause and the embedded clause are question-marked: "You `know-not-know` tomorrow Boston `will-not-will` snow?" The particle form of the question obligatorily has only one ma5 at the end of the sentence: "You know tomorrow Boston will snow `ma5`?" As previously noted, the locative and temporal expressions, "Boston" and "tomorrow" are typically brought to the front of their clause constituent for both questions and statements in Mandarin.

We have solved the understanding half of the previous problem by employing the trace mechanism in parsing to produce a homogeneous semantic frame. Similarly, the generation system needs to rearrange the hierarchical structure of the semantic frame to achieve the desired surface ordering in the target language. Our generation system supports two mechanisms for dealing with movement which we call a push and a pull mechanism, respectively. The push mechanism decouples the generation of a constituent from its placement in the surface string, allowing a constituent which needs to be preposed to be generated at its deep-structure location and placed typically at the head of its parent clause. The pull mechanism allows the parent clause to control both generation and placement of a deeply-nested constituent.

The handling of question form in Mandarin is intricate and involves both the lexicon and the preprocessor in addition to the main generation processor. It is rendered more complex in part because we wanted to be able to allow both forms of questions to be produced as alternates. In the preprocessing stage, an additional feature is introduced into the semantic frame whenever the clause is a yes-no question. When the feature is present, the lexical entries for several verbs, auxiliaries, and adjectives, for example, "know," "will," and "cold," cycle through two alternatives, one introducing the A-not-A question form and the other introducing the q-particle based question form. In order to prevent a duplicate ma5 for the parent and subordinate clauses in sentences of the form, " do you know if ...", a mechanism in the main processor is able to suppress the second ma5 on the basis of a flag instantiated during production of the first one.

Another important aspect of high quality translation is the selection of the appropriate word sense for multisense words. A simple example that is prevalent in the weather domain is the multiple interpretations of the word "what." Consider the two seemingly similar sentences "What is the weather tomorrow?" and "What is the temperature tomorrow?" For "weather" the appropriate translation for "what" is "zen3 me5 yang4" (literally "how"), whereas for "temperature," a correct translation is "duo1 shao3" which would literally mean "how much." Our generation framework allows the NP constituent to set up a semantic flag which is then available as a conditioning context for selecting the appropriate sense of the word "what." Our lexical entry for "what" includes five distinct conditioning contexts.

5. EXAMPLE-BASED TRANSLATION

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

It is natural in our system to index the translation table using some form of interlingua. In a way, our example-based translation follows the same paradigm (i.e., interlingua based translation) as the formal method: parsing and generation. The only difference is that the target sentence is generated by looking up pre-existing examples using an interlingua as the matching criterion. The complexity of the index determines the degree of correspondence between the matched translation pairs: a more detailed index potentially leads to a closer match but at the increased risk of search failure due to data sparseness. We

```
{c eform
   :WEATHER "rain"
   :CITY "<CITY>"
   :DATE "<DATE>"
   :CLAUSE "verify"
   :sentences (
     "<CITY> <DATE> hui4 xia4 yu3 ma5?"
     "<CITY> <DATE> hui4 bu2 hui4 xia4 yu3?"
     "<CITY> <DATE> xia4 bu2 xia4 yu3?"
     "<DATE> <CITY> hui4 xia4 yu3 ma5?"
     "ni3 zhi1 dao4 <CITY> <DATE> xia4 yu3 ma5?"
     ... ) }
```

Fig. 6. Example of a group of Chinese sentences with the same key-value index.

use very lean syntactic and semantic information, encoded as key-value pairs, to index our automatically generated translation corpus. In the following, we describe how to generate an indexed corpus as well as the retrieval mechanism to search the corpus.

5.1 Generation of Translation Corpus

Our example-based translation begins with the construction of an indexed translation corpus which is done automatically by utilizing the parsing and generation components, along with a monolingual English corpus. Each English sentence is first parsed to yield a semantic frame from which a Chinese translation is derived using formal generation rules as described in Section 2 and 3. In addition, a set of trivial generation rules are created to extract very lean semantic and syntactic information from the semantic frame as key-value pairs which can then be used as an index for the Chinese sentence. For example, from the semantic frame shown in Figure 5, the Chinese translation would be “bo1 shi4 dun4 zhei4 zhou1 mo4 hui4 xia4 yu3 ma5?” (among other variations), and the KV representation is “WEATHER: rain CITY: Boston DATE: this weekend CLAUSE: verify.”

In order to prevent erroneous translations from confusing the student (who is trying to learn the language), we use the L2 grammar developed for the dialogue system to ensure that the translation output is a legitimate sentence. We noticed that the grammar would occasionally reject good translations due to coverage gaps in the L2 grammar. However, we think that this is a desirable feature because those sentences will nevertheless fail in subsequent system processing if the students choose to imitate them during their conversation with the system.

To improve efficiency, all unique Chinese sentences with the same set of key-value pairs (ignoring order) are grouped together in the indexed corpus. We can further reduce the corpus size (and data sparseness) by mapping the values of certain keys (e.g., city names, months, dates, etc.) to a generic class label, both in the key-value index as well as in the sentence. We refer to these keys as *classed keys*. Figure 6 shows a typical group of such indexed sentences/templates. Given the thin KV index, it is possible to have sentences with very different surface forms map to the same index. This is a useful feature for language learning: we can present multiple translation choices to a student for increased exposure to variability.

5.2 Retrieval Mechanism

The basic function of the retrieval mechanism is to find a candidate sentence whose KV index matches the input KV specification. To allow certain flexibility in matching the key-value pairs and alleviate the data sparseness problem, keys are differentiated into several categories depending on whether they are optional or obligatory and whether or not the value of a key is masked during the matching. These are specified in a header file in the indexed corpus in order to allow a developer to flexibly modify the matching strategy. Each obligatory key in the input KV specification has to be accounted for in the matching process, while optional keys in the input can be ignored to avoid a matching failure (but will be preferred otherwise). Value masking during retrieval typically applies to the classed keys (e.g., city names). After a sentence template is retrieved, the values of class keys are first translated by simple lexical lookup ("Boston" becomes "bo1 shi4 dun4") and reinserted into the appropriate slots in the surface string. This is equivalent to the technique of replacing lexical entries with classes during example-based matching described in Brown [1999]. If more than one group of sentences is retrieved, the selection pool includes all the groups.

We will illustrate the retrieval process with an example to highlight some of the distinctions in the different key types. Assume we want to retrieve from the indexed corpus a Chinese translation for "What is the chance of rain in Seattle tomorrow?" The parsing and generation systems will first produce the following key-value pairs for the input.

WEATHER:	rain
CITY:	Seattle
DATE:	tomorrow
CLAUSE:	wh-question

Suppose the corpus contains only the example shown in Figure 6, with WEATHER as an obligatory key required to match on the key-value level, CITY and DATE obligatory keys required to match on the key level, while CLAUSE is an optional key required to match on the key-value level. All of the sentence templates are legitimate candidates, and a possible translation would be:

xi1 ya3 tu2	ming2 tian1	hui4	xia4 yu3	ma5
Seattle	*tomorrow*	*will*	*rain*	*question-particle*

If the CLAUSE were specified as an obligatory key matching on the key-value level, then the search would fail to generate any output. For an input such as "Will it rain in Seattle?", the search would also fail because of the extra obligatory key DATE in the corpus.

Notice that the example-based translation does not necessarily convey the exact meaning as the input because the KV information can be lossy (e.g.,

Multiple sentences
Hi, I am in Boston. What is the weather like today?
`CITY: Boston WEATHER: weather DATE: today`

False starts
How hot is it - How hot is it in Seattle?
`CITY: Seattle WEATHER: hot`

Recognition error
I am in Boston is going to rain today
`CITY: Boston WEATHER: rain DATE: today`

Fig. 7. Examples of problematic sentences with associated KV strings in development data.

the nuance of “chance of rain” is lost in the translation). However, since KV can usually be extracted even with keyword spotting, this method is much more robust to unexpected linguistic patterns and ill-formed sentences. Figure 7 illustrates some example sentences observed in our development data with associated KV strings.

6. EVALUATION

We conducted evaluation experiments in the weather query domain using English spontaneous speech data recorded from telephone calls to the publicly-available JUPITER weather information system [Zue et al. 2000]. Our test data consists of 695 utterances selected from a set of held-out data previously defined to facilitate speech recognition evaluations. Utterances whose manually-derived transcription can not be parsed by the English grammar are excluded from the evaluation since they are likely to be out-of-domain sentences and would simply contribute to null outputs. Metalevel sentences such as “start over” and “goodbye” are also excluded since our system in its usage responds to them rather than translating them. The test data have on average 6.5 words per utterance. The recognizer uses the MIT landmark-based SUMMIT system [Glass 2003] which adopts a finite state transducer-based search algorithm. The class n-gram language model is automatically generated from the natural language grammar using techniques described in Seneff et al. [2003]. The recognizer achieved 6.9% word error rate and 19.0% sentence error rate on this set.

In addition to the held-out data, we also have over 120,000 transcribed utterances collected via the JUPITER system used for training the recognizer’s acoustic and language models. These data enabled us to derive a large indexed translation corpus for the example-based translation system. We also used these data as a development set to improve coverage of the formal rule-based system.

The availability of this large quantity of domain-specific data also allowed us to explore other machine translation methods, such as statistical-based translation, which typically requires large amounts of bi-text data for adequate training. We use the formal translation procedure to automatically obtain a parallel English-Chinese bilingual corpus from the orthographic transcriptions. We again use parsability by the Chinese grammar as a quality check of the generated translation which yielded about 50,000 high-quality parallel translation pairs.

Among statistical machine translation systems, phrase-based methods have been shown to produce the best translation performance [Koehn 2004]. For the baseline system, we trained a state-of-the-art phrase-based model [Koehn et al. 2003; Koehn 2004] using the 50K English-Chinese sentence pairs in the weather domain. The training process involves deriving word alignments using GIZA++ [Och and Ney 2003] and extracting and scoring translation phrases from the word-aligned data. During testing, we simply used the default settings of the Pharaoh decoder [Koehn 2004] without special parameter tuning.

The translation quality was manually rated by a bilingual judge based on grammaticality and fidelity where the input to the translation module is either the manual transcription of the test utterances or the speech recognition outputs (*1*-best or *10*-best lists). In all cases, fidelity is judged against the reference orthographies so that errors caused by the speech recognition system (e.g., a misrecognized date or city) result in translation errors. We designed a highly efficient procedure to minimize the amount of required manual ratings as well as to reduce human inconsistency and bias similar to the methodology adopted in Ney et al. [2000].

The rating procedure works as follow. First, the translations for the test utterances are generated under each test condition. The translations are then grouped by the corresponding reference orthographies. In our experiments, we generated 2,207 unique translations which are grouped by 599 unique reference orthography for the 695 test utterances. A bilingual judge then rated the 599 groups of translations without explicit knowledge of the testing condition of each translation, which avoids bias towards any testing condition. The rating procedure enables the judge to rate each unique translation pair once and only once. The judge can also compare the translations within a group when rating the quality of each translation which leads to improved consistency of the ratings. Three categories are used in the manual rating: `Perfect (P)`, `Acceptable (A)`, and `Incorrect (I)`. It is possible for our system to produce null output which is categorized as `Failed (F)`. Figure 8 shows some example translation outputs and the corresponding human ratings.

Two criteria are defined to evaluate the translation performance: accuracy and success rate. *Accuracy* is defined as the percentage of perfect and acceptable outputs over the total number of nonempty outputs, that is,

$$Accuracy = \frac{P + A}{P + A + I}. \tag{1}$$

Success Rate is defined as the percentage of perfect and acceptable outputs over all data:

$$SuccessRate = \frac{P + A}{P + A + I + F}. \tag{2}$$

In the following, we first compare the performance of the three individual translation methods. We then report performance of the integrated system in our language tutoring system. Some analysis of the translation outputs is also provided.

1. Can you give me a forecast for tomorrow?
1a. ni3 neng2 bu4 neng2 gei3 wo3 ming2 tian1 de5 tian1 qi4 P
1b. wo3 neng2 you3 ming2 tian1 de5 tian1 qi4 yu4 bao4 ma5 P
1c. ni3 you3 ming2 tian1 de5 xin4 xi1 ma5 P
1d. ni3 you3 ming2 tian1 de5 tian1 qi4 yu4 bao4 ma5 P
1e. ni3 neng2 gei3 wo3 ming2 tian1 de5 tian1 qi4 ma5 P

2. What are the chances that it is going to rain in Boston tonight?
2a. bo1 shi4 dun4 jin1 wan3 hui4 bu2 hui4 xia4 ji1 luu4 shi4 duo1 shao3 yi2 xia4 ma5 I
2b. bo1 shi4 dun4 jin1 wan3 de5 xia4 yu3 ji1 luu4 shi4 duo1 shao3 P
2c. jin1 wan3 bo1 shi4 dun4 hui4 xia4 yu3 de5 ji1 luu4 shi4 duo1 shao3 P
2d. bo1 shi4 dun4 jin1 wan3 hui4 xia4 yu3 de5 ji1 luu4 shi4 duo1 shao3 P
2e. bo1 shi4 dun4 jin1 wan3 hui4 xia4 yu3 ma5 A

3. Can we make snow in Boston?
3a. ni3 neng2 wo3 men2 make bo1 shi4 dun4 xia4 xue3 ma5 I
3b. bo1 shi4 dun4 you3 mei2 you3 xue3 I
3c. wo3 men2 neng2 bu4 neng2 I
3d. bo1 shi4 dun4 you3 xue3 ma5 I
3e. wo3 men2 neng2 ma5 I
3f. bo1 shi4 dun4 hui4 bu2 hui4 xia4 xue3 I
3g. bo1 shi4 dun4 xia4 xue3 ma5 I
3h. bo1 shi4 dun4 you3 bao4 feng1 xue3 ma5 I
3i. bo1 shi4 dun4 hui4 xia4 xue3 ma5 I

Fig. 8. Translation outputs and human ratings for three English inputs. (P = Perfect, A = Acceptable, I = Incorrect)

Table I. Performance of Each Translation Method in Text Mode on a Set of 695 Unseen Test Utterances

Quality	P + A	I	F	Accuracy	Success Rate
Formal	633 + 27	31	4	95.5%	95.0%
Example	607 + 45	23	20	96.6%	93.8%
Statistical	551 + 43	101	0	85.5%	85.5%

Table II. Performance of Each Translation Method on Speech Recognition Outputs (*1*-Best or *10*-Best) on a Set of 695 Utterances

Quality	P + A	I	F	Accuracy	Success Rate
Formal (*1*-best)	585 + 31	55	24	91.8%	88.6%
Formal (*10*-best)	599 + 29	57	10	91.7%	90.4%
Example (*1*-best)	564 + 53	37	41	94.3%	88.8%
Example (*10*-best)	575 + 50	42	28	93.7%	89.9%
Statistical (*1*-best)	516 + 47	127	5	81.6%	81.0%

6.1 Performance of Individual Translation Methods

Table I and Table II summarize the performance of each translation method in text mode and speech mode, respectively. We do not impose a parsability check in this configuration; however, the example-based method has an implicit parsability check due to the way in which the translation corpus is generated. The formal method achieves very good performance in text mode (95% overall success rate). The example-based method alone did not outperform the formal method in overall success rate, despite the fact that it is expected to be more robust. This is because the translation corpus only represents a subspace of the outputs reachable by the generation method (due to the parsability check as well as data sparseness) which resulted in more failures. We also suspect that most of the text inputs are well-formed which gives the formal method an advantage. The statistical method is always able to generate an output and achieved an 85.5% success rate in text mode.

Not surprisingly, the performances degraded when moving to speech mode for all methods. As expected, the degradation is smaller for the example-based method than for the formal method. The example-based method has a comparable success rate to the formal method, while the accuracy is higher (due to the implicit parsability check). Since our natural language parsing system can parse *N*-best lists, we also explored using a *10*-best list as the interface between the recognizer and the translation system as shown in Table II. The *10*-best list improved the overall success rate, although with a slight degradation in accuracy for both the formal and example-based methods because the natural language understanding system is able to parse more utterances with a deeper search space.

It is interesting to observe some of the characteristics of the outputs of different translation methods. For example, Sentence 2 in Figure 8, "What are the chances that it is going to rain in Boston tonight?" represents a challenging sentence that has long-distance dependencies and that was unobserved in the training data. One of the perfect Chinese translations (2c in Figure 8) produced by the formal method is:

jin1 wan3	bo1 shi4 dun4	hui4 xia4 yu3	de5	ji1 luu4	shi4 duo1 shao3
tonight	Boston	will rain	+s	chance	is how much

In comparison, the output from the statistical method (2a) is a string of locally-(almost) coherent short phrases but does not correspond to an overall well-formed Chinese sentence.

bo1 shi4 dun4	jin1 wan3	hui4 bu2 hui4	xia4	ji1 luu4
Boston	tonight	will-not-will	fall	chance
shi4 duo1 shao3	yi2 xia4	ma5		
is how much	once	question-particle		

Notice that the incompatibility of "will-not-will" and "question-particle" is also violated.

The output from the example-based method (2e) received an acceptable rating because the notion of chance is missing from the otherwise well-formed translation output:

bo1 shi4 dun4	jin1 wan3	hui4	xia4 yu3	ma5
Boston	tonight	will	rain	question-particle

This is because the KV string from the input, "`CLAUSE: verify CITY: Boston Date: tonight WEATHER: rain`," is simply mapped to a Chinese sentence equivalent to "Will it rain in Boston tonight?" during example retrieval.

Table III. Performance of Formal, Example, and Formal + Example Methods in Text Mode on a Set of 695 Unseen Test Utterances (All methods include a parsability check.)

Quality	P + A	I	F	Accuracy	Success Rate
Formal	612 + 9	6	68	98.9%	89.4%
Example	607 + 45	23	20	96.6%	93.8%
Formal + Example	644 + 17	21	13	96.9%	95.1%

Table IV. Performance of Formal, Example, and Formal + Example Methods on Automatic Speech Recognition Outputs (*10*-best list) on a Set of 695 Unseen Test Utterances (All methods include a parsability check.)

Quality	P + A	I	F	Accuracy	Success Rate
Formal	582 + 17	19	77	96.9%	86.2%
Example	575 + 50	42	28	93.7%	89.9%
Formal + Example	610 + 21	43	21	93.6%	90.8%

Sentence 3 in Figure 8 represents an out-of-domain query for which all methods produced incorrect translations, however, in very different ways. The statistical method simply strung together word/phrase translations (with the word "make" untranslated) which resulted in "you can we make Boston snow question-particle" (output 3a). The generation method produced the Chinese equivalent of "can we" (3c and 3e) because the predicate "make" is absent from the generation rules. The example-based method produced the equivalent of "Will it snow in Boston?" (3b,3d,3f-i) because the KV string for the input is simply "`CLAUSE: verify CITY: Boston WEATHER: snow`."

6.2 Performance of Integrated System

In our deployed language tutoring system, we adopt the strategy of preferring the formal generation-based output if it can be accepted by the Chinese grammar. The formal method is able to achieve high fidelity in the translation, preserving syntactic correspondence between English and Chinese as much as possible. A parsability check is imposed on the formal method, and we back off to the example-based method if the generation method failed.

Table III and Table IV summarize the performance of the combined formal + example method with comparison with each individual method (with parsability check), in text mode and speech mode, respectively. Notice that the accuracy of the formal method increased to almost 99%, along with the drop in success rate down to below 90%. This drop is partially due to coverage gaps in the Chinese grammar used for checking. The example-based method is unaffected because of the implicit checking already in place in the translation corpus, and out-performs the formal method in this configuration. However, the combined method is able to improve both the success rate and accuracy over the example-based method alone in text mode. Similar trends can be observed in speech mode, although the performances of all methods have degraded, as expected.

6.3 Error Analysis

We performed a detailed analysis of the errors made by the integrated system (formal + example) in speech mode. We observed that 18 of the 43 incorrect translations (41.9%) are caused by speech recognition errors (insertions, deletions, and substitutions) on city names and dates. An additional 11 incorrect translations (25.6%) are caused by recognition errors on what type of information the user is seeking (weather, high/low temperature, etc.). Since our language learning system paraphrases the recognized user query and plays it back to the user, these types of errors are unlikely to cause confusions because the user is aware of the context of the translation error.

There are 11 errors (25.6%) caused by gaps in domain coverage which include (arguably) 7 out-of-domain queries, such as:

—Can we make snow in Boston?
—What is a tornado?
—Should I go to work tomorrow morning?
—Where is your location?

We hope that these types of queries will occur minimally among language learners with adequate user education and lesson preparation.

The remaining 3 incorrect translations are caused by errors/inadequacies in the natural language understanding and generation components. For example, "What is the weather going to be like in Boston *tomorrow Tuesday*?" was interpreted as "*tomorrow and Tuesday*" by the parser. Another query "Do you *only* know information about weather?" was translated into the Chinese equivalent of "Do you know weather information?" because the adverb "*only*" was missed in generation. These errors are more harmful to the language learners, but, fortunately, they account for less than 7% of the errors in our system. Furthermore, once identified, they can easily be corrected in the rules used by the parsing and generation components.

7. RELATED WORK

Spoken language translation first emerged as an attractive, if challenging, research topic in the mid 1980's, driven in part by the ambitious goals set forth by the then newly, formed Advanced Technology Research (ATR) laboratories in Japan. ATR played a leadership role in advocating the example-based translation method (EBMT) [Sumita and Iida 1991; Sato 1992; Iida et al. 1996; Sumita 2001; Shimohata et al. 2003] which is particularly well suited to languages with widely differing word order such as English and Japanese. While initial efforts involved extensive manual annotation of training data, techniques for utterance tagging eventually became more automatic, and indexing in the original EBMT has migrated towards a new paradigm called TDMT (transfer-driven MT) where examples are decomposed into phrasal units and a second stage assembles the retrieved phrases into well-formed sentences [Seligman 2000]. While our own approach to EBMT does exploit classes for generalization, we have not yet attempted to break down sentences into phrasal units for increased generality from our example templates.

The Multilingual Automatic Translation System (MARS) at IBM [Gao et al. 2002] is a project aimed at conversational speech-to-speech translation between English and Mandarin Chinese for limited domains. Their approach is similar to ours in that translation is modeled as a cascade process of parsing and generation. However, the MARS system uses only semantic information in the interlingua and adopts a statistical approach to parsing and natural language generation. The semantic parser and natural language generation components are both trained from annotated domain-specific data, thus no handcrafted grammars or rules are needed. In contrast, our interlingua representation captures both syntactic and semantic information and the parsing and generation adopt a formal framework.

The system in the literature that most resembles our formal method is probably the Janus system [Levin et al. 2000]. In fact, over the years, their strategies have exhibited marked parallels to ours even when we were mainly concerned with multilingual dialogue systems. They, like us, utilized semantic rules to encode domain-specific knowledge and advocated parsing from multiple languages into a common meaning representation. Our multilingual systems always included a paraphrase of the user query that was displayed or spoken back to the user to confirm understanding and help maintain dialogue coherence in the face of possible recognition errors. Likewise, their translation system offered a paraphrase in the source language to the speaker for verification. Furthermore, both systems have focused on the travel domain, and both have been working towards the goal of merging grammars from multiple domains into a single shared grammar resource with common elements such as dates and times compartmentalized into subgrammars. We have taken this exercise a step further in capturing syntactic structure in the upper nodes of the parse tree, thus providing a more detailed accounting of the syntax and easing the pain of grammar development for new database query domains.

Spearheaded by the ambitious efforts of IBM [Brown et al. 1990, 1993], statistical machine translation emerged in the early 1990's as a powerful technique for handling the many ambiguities in language without requiring painstaking effort from linguists. For text translation, statistical methods are viewed today as offering clear advantages, particularly when bilingual corpora can be discovered on the Web and sentence alignments can be largely automated [Och and Ney 2003]. Most statistical translation methods decompose the problem into three components: a lexical model, an alignment model, and a target language model [Ney et al. 2000]. Phrase-based statistical MT, which is relatively new [Wang and Waibel 1998; Och et al. 1999; Alshawi et al. 2000; Zens et al. 2002; Koehn et al. 2003], improves word-based alignment models by exploiting shallow phrase structures. Phrase-based techniques have been quite competitive in recent evaluations [Koehn et al. 2003].

8. CONCLUSIONS AND FUTURE WORK

In this article, we have demonstrated an effective approach for domain-restricted speech-to-speech translation from English to Mandarin Chinese, based on a combination of formal and example-based translation techniques.

Evaluated on an unseen test set of nearly 700 utterances, we have shown that the system can outperform a state-of-the-art statistical method. We believe that formal methods are appropriate for the situation where translation quality needs to be extremely high but the domain is restricted. Statistical methods are limited by the localized patterns observed in the training data and are therefore sometimes unable to handle long-distance dependencies appropriately. Using formal methods to create a translation corpus for the example-based approach is an effective strategy for handling ill-formed sentences, consequential to either spontaneous speech disfluencies or recognition errors.

While we believe that our methodology will be effective for language learning applications, we have yet to demonstrate that this is the case. We also want to port it to many other applications besides weather to form a suite of lesson plans on different topics. Our syntax-based formulation of the English grammar will ease the burden of porting to other domains. We are also conducting research on automatic techniques to induce the Chinese grammar given the English-to-Chinese translation framework [Lee and Seneff 2004].

ACKNOWLEDGMENTS

The authors would like to thank Dr. Philipp Koehn for his generous help with the Pharaoh decoder for phrase-based statistical machine translation models.

REFERENCES

ALSHAWI, H., BANGALORE, S., AND DOUGLAS, S. 2000. Head-transducer models for speech translation and their automatic acquisition from bilingual data. *Machine Transla. 15*, 105–124.

BANGALORE, S. AND RICCARDI, G. 2002. Stochastic finite-state models for spoken language machine translation. *Machine Transla. 17*, 165–184.

BAPTIST, L. AND SENEFF, S. 2000. Genesis-II: A versatile system for language generation in conversational system applications. In *Proceedings of the International Conference on Spoken Language Processing*. Beijing, China.

BROWN, P. F., COCKE, J., PIETRA, S. A. D., PIETRA, V. J. D., JELINEK, F., MERCER, R. L., AND ROOSSIN, P. S. 1990. A statistical approach to machine translation. *Computat. Linguist. 16*, 2, 79–85.

BROWN, P. F., PIETRA, S. A. D., PIETRA, V. J. D., AND MERCER, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computat. Linguist. 19*, 2, 263–311.

BROWN, R. D. 1999. Adding linguistic knowledge to a lexical example-based translation system. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*. Chester, England.

CASACUBERTA, F., NEY, H., OCH, F. J., VIDAL, E., VILAR, J. M., BARRACHINA, S., GARCIA-VAREA, I., LLORENS, D., MARTINEZ, C., MOLAU, S., NEVADO, F., PASTOR, M., PICO, D., SANCHIS, A., AND TILLMANN, C. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Comput. Speech Lang. 18*, 24–47.

COLLINS, M. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the Association for Computational Linguistics*. Madrid, Spain.

COWAN, B. 2004. PLUTO: A preprocessor for multilingual spoken language generation. M.S. thesis, MIT, Cambridge, MA.

FREDERKING, R. E., BLACK, A. W., BROWN, R. D., RUDNICKY, A., MOODY, J., AND STEINBRECHER, E. 2002. Speech translation on a tight budget without enough data. In *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*. Philadelphia, PA.

GAO, Y., ZHOU, B., DIAO, Z., SORENSEN, J., ERDOGAN, H., AND SARIKAYA, R. 2002. A trainable approach for multilingual speech-to-speech translation system. In *Proceedings of the Human Language Technology Conference*. San Diego, CA.

GAO, Y., ZHOU, B., DIAO, Z., SORENSEN, J., AND PICHENY, M. 2002. MARS: a statistical sematnic parsing and generation-based multilingual automatic translation system. *Machine Translat. 17*, 185–212.

GLASS, J. 2003. A probabilistic framework for segment-based speech recognition. *Comput. Speech Lang. 17*, 137–152.

GODFREY, J., HOLLIMAN, E., AND MCDANIEL, J. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. San Francisco, CA, 517–520.

HE, Y. AND YOUNG, S. 2003. A data-driven spoken language understanding system. In *Proceedings of Automatic Speech Recognition and Understanding*. St Thomas, US Virgin Islands.

IIDA, H., SUMITA, E., AND FURUSE, O. 1996. Spoken-language translation method using examples. In *Proceedings of the 16th Conference on Computational Linguistics*. Vol. 2. Copenhagen, Denmark, 1074–1077.

KOEHN, P. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Association for Machine Translation in the Americas*. Washington DC.

KOEHN, P., OCH, F. J., AND MARCU, D. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference-North American Chapter of the Association for Computational Linguistics*. Edmonton, Canada.

KUMAN, S. AND BYRNE, W. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of the Human Language Technology Conference-North American Chapter of the Association for Computational Linguistics*. Edmonton, Canada.

LEE, J. AND SENEFF, S. 2004. Translingual grammar induction. In *Proceedings of the International Conference on Spoken Language Processing*. Jeju Island, Korea.

LEVIN, L., LAVIE, A., WOSZCZYNA, M., GATES, D., GAVALDA, M., KOLL, D., AND WAIBEL, A. 2000. The Janus III translation system: speech-to-speech translation in multiple domains. *Machine Transla. 15*, 1-2, 3–25. Special Issue on Spoken Language Transla.

MARCU, D. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of the Association for Computational Linguistics*. Toulouse, France.

MARCUS, M. P., SANTORINI, B., AND MARCINKIEWICZ, M. A. 1993. Building a large annotated corpus of English: the Penn TreeBank. *Computat. Linguist. 19*, 2, 313–330.

NEY, H., NIESSEN, S., OCH, F. J., SAWAF, H., TILLMANN, C., AND VOGEL, S. 2000. Algorithms for statistical translation of spoken language. *IEEE Trans. Speech Audio Proces. 8*, 1, 24–36.

OCH, F. J. AND NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computat. Linguist. 29*, 1, 19–52.

OCH, F. J., TILLMANN, C., AND NEY, H. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. College Park, MD, 20–28.

PRICE, P. 1990. Evaluation of spoken language systems: the ATIS domain. In *Proceedings of the DARPA Speech and Natural Language Workshop*. Hidden Valley, PA, 91–95.

RAYNER, M. AND CARTER, S. 1997. Hybrid processing in the spoken language translator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Munich, Germany.

SATO, S. 1992. CTM: An example-based translation aid system. In *Proceedings of the International Conference on Computational Linguistics*. Nantes, France.

SELIGMAN, M. 2000. Nine issues in speech translation. *Machine Translation 15*, 149–185.

SENEFF, S. 1992a. Robust parsing for spoken language systems. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. San Francisco, CA, 189–192.

SENEFF, S. 1992b. TINA: A natural language system for spoken language applications. *Computat. Linguist. 18*, 1, 711–714.

SENEFF, S. AND POLIFRONI, J. 2000. Dialogue management in the MERCURY flight reservation system. In *Proceedings of the Applied Natural Language Conference, Satellite Workshop*. Seattle, WA.

SENEFF, S., WANG, C., AND HAZEN, T. J. 2003. Automatic induction of *n*-gram language models from a natural language grammar. In *Proceedings of the Eurospeech*. Geneva, Switzerland.

SENEFF, S., WANG, C., AND ZHANG, J. 2004. Spoken conversational interaction for language learning. In *Proceedings of National Language Processing and Speech Technologies in Advanced Language Learning Systems*. Venice, Italy.

SHIMOHATA, M., SUMITA, E., AND MATSUMOTO, Y. 2003. Retrieving meaning-equivalent sentences for example-based rough translation. In *Proceedings of the Human Language Technology*. Edmonton, Canada, 50–56.

SUMITA, E. 2001. Example-based machine translation using dp-matching between word sequences. *Workshop on Data-Driven Methods in Machine Translation*. Toulouse, France, 1–8.

SUMITA, E. AND IIDA, H. 1991. Experiments and prospects of example-based machine translation. In *Proceedings of the Association for Computational Linguistics*. Berkley, CA, 185–192.

TANG, M., LUO, X., AND ROUKOS, S. 2002. Active learning for statistical natural language parsing. In *Proceedings of the Association for Computational Linguistics*. Philadelphia, PA.

VEALE, T. AND WAY, A. 1997. Gaijin: A template-driven bootstrapping approach to example-based machine translation. In *Proceedings of the NMNLP*. Sofia, Bulgaria.

WANG, C. AND SENEFF, S. 2004. High-quality speech translation for language learning. In *Proceedings of Natural Language Processing: National Language Processing and Speech Technologies in Advanced Language Learning Systems*. Venice, Italy.

WANG, Y. AND WAIBEL, A. 1998. Modeling with structures in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*. Montreal, Quebec, Canada, 1357–1363.

ZENS, R., OCH, F. J., AND NEY, H. 2002. Phrase-based statistical machine translation. In *Proceedings of the 25th German Conference on Artificial Intelligence*. Aachen, Germany, 18–32.

ZUE, V., SENEFF, S., GLASS, J., POLIFRONI, J., PAO, C., HAZEN, T. J., AND HETHERINGTON, L. 2000. JUPITER: A telephone-based conversational interface for weather information. *IEEE Trans. Speech Audio Process. 8*, 1, 85–96.

Received August 2005; accepted March 2006 by Kishore Papineni