ABSTRACT

At MIT Lincoln Laboratory, we have been developing a Korean-to-English machine translation system CCLINC (Common Coalition Language System at Lincoln Laboratory). The CCLINC Korean-to-English translation system consists of two core modules, language understanding and generation modules mediated by a language neutral meaning representation called a semantic frame. The key features of the system include: (i) Robust efficient parsing of Korean (a verb final language with overt case markers, relatively free word order, and frequent omissions of arguments). (ii) High quality translation via word sense disambiguation and accurate word order generation of the target language. (iii) Rapid system development and porting to new domains via knowledge-based automated acquisition of grammars. Having been trained on Korean newspaper articles on “missiles” and “chemical biological warfare,” the system produces the translation output sufficient for content understanding of the original document.

1. SYSTEM OVERVIEW

The CCLINC Korean-to-English translation system is a component of the CCLINC Translingual Information System, the focus languages of which are English and Korean, [11,17]. Translingual Information System Structure is given in Figure 1.

Given the input text or speech, the language understanding system parses the input, and transforms the parsing output into a language neutral meaning representation called a semantic frame, [16,17]. The semantic frame — the key properties of which will be discussed in Section 2.3 — becomes the input to the generation system. The generation system produces the target to the generation system, the semantic frame can be utilized for other applications such as translingual information extraction and language translation output after word order arrangement, vocabulary replacement, and the appropriate surface form realization in the target language, [6]. Besides serving as the input question-answering, [12]. In this paper, we focus on the Korean-to-English text translation component of CCLINC.

2. ROBUST PARSING, MEANING REPRESENTATION, AND AUTOMATED GRAMMAR ACQUISITION

This work was sponsored by the Defense Advanced Research Project Agency under the contract number F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

For other approaches to Korean-to-English translation, the readers are referred to *Korean-to-English translation* by Egedi, Palmer, Park and Joshi 1994, a transfer-based approach using synchronous tree adjoining grammar, [5], and Dorr 1997, a small-scale interlingua-based approach, using Jackendoff’s lexical conceptual structure as the interlingua, [4].
Interlingua-Based Board-Coverage Korean-to-English Translation in CCLINC

Massachusetts Institute of Technology, Laboratory for Computer Science, Spoken Language Systems Group, Cambridge, MA, 02139

Approved for public release; distribution unlimited

The original document contains color images.
1.1 Robust Parsing

The CCLINC parsing module, TINA [16], implements the top-down chart parsing and the best-first search techniques, driven by context free grammars rules compiled into a recursive transition network augmented by features, [8]. The following properties of Korean induce a great degree of ambiguity in the grammar: (i) relatively free word order for arguments —— given a sentence with three arguments, subject, object, indirect object, all 6 logical word order permutations are possible in reality, (ii) frequent omissions of subjects and objects, and (iii) the strict verb finality, [10]. Due to the free word order and argument omissions, the first word of an input sentence can be many way ambiguous —— it can be a part of a subject, an object, and any other post-positional phrases. The ambiguity introduced by the first input word grows rapidly as the parser processes subsequent input words. Verbs, which usually play a crucial role in reducing the ambiguity in English by the subcategorization frame information, are not available until the end, [1,3,11].

Our solution to the ambiguity problem lies in a novel grammar writing technique, which reduces the ambiguity of the first input word. We hypothesize that (i) the initial symbol in the grammar (i.e. Sentence) always starts with the single category generic_np, the grammatical function (subject, object) of which is undetermined. This ensures that the ambiguity of the first input word is reduced to the number of different ways the category generic_np can be rewritten. (ii) The grammatical function of the generic_np is determined after the parser processes the following case marker via a trace mechanism.

Figure 2 illustrates a set of sample context free grammar rules, and Figure 3 (on the next page) is a sample parse tree for the input sentence “URi Ga EoRyeoUn MunJe Reul PulEox Da (We solved a difficult problem).”

(i) sentence → generic_np clause sentence_marker
(ii) clause → subject generic_np object verbs
(iii) subject → subj_marker np_trace

Figure 2. Sample context free grammar rules for Korean

The generic_np dominated by the initial symbol sentence in (i) of Figure 2 is parsed as an element moved from the position occupied by np_trace in (iii), and therefore corresponds to the category np_trace dominated by subject in Figure 3 (placed on the next page for space reasons). All of the subsequent generic_np’s, which are a part of a direct object, an indirect object, a post-positional phrase, etc. are unitarily handled by the same trace mechanism. By hypothesizing that all sentences start with generic_np, the system can parse Korean robustly and efficiently. The trace mechanism determines the grammatical function of generic_np by repositioning it after the appropriate case marker.

Utilization of overt case markers to improve the parsing efficiency precisely captures the commonly shared intuition for parsing relatively free word order languages with overt case markers such as Korean and Japanese, compared with parsing relatively strict word order languages with no overt case markers such as English: In languages like English, the verb of a sentence plays the crucial role in reducing the ambiguity via the verb subcategorization frame information on the co-occurring noun phrases, [1,3,11]. In languages like Korean, however, it is typically the case marker which identifies the grammatical function of the co-occurring noun phrase, assuming the role similar to that of verbs in English. The current proposal is the first explicit implementation of this intuition, instantiated by the novel idea that all noun phrases are moved out of the case marked phrases immediately following them.

2.2 Meaning Representation and Generation

The CCLINC Korean-to-English translation system achieves high quality translation by (i) robust mapping of the parsing output into the semantic frame, and (ii) word sense disambiguation on the basis of the selection preference between two grammatical relations (verb-object, subject-verb, head-modifier) easily identifiable from the semantic frame, [13]. The former facilitates the accurate word order generation of various target language sentences, and the latter, the accurate choice of the target language word given multiple translation candidates for the same source language word. Given the parsing output in Figure 3, the system produces the semantic frame in Figure 4.

2 Post-positional phrases in Korean correspond to pre-positional phrases in English. We use the term post-positional phrase to indicate that the function words at issue are located after the head noun.

3 The hypothesis that all sentences start with a single category generic_np is clearly over simplified. We can easily find a sentence starting with other elements such as coordination markers which do not fall under generic_np. For the sentences which do not start with the category generic_np, we discard these elements for parsing purposes. And this method has proven to be quite effective in the overall design of the translation system, especially due to the fact that most of non generic_np sentence initial elements (e.g. coordination markers, adverbs, etc.) do not contribute to the core meaning of the input sentence.

4 Throughout this paper, “subj_marker” stands for “subject marker”, and “obj_marker”, “object marker”.

5 Strictly speaking, the meaning representation in Figure 4 is not truly language neutral in that the terminal vocabularies are represented in Korean rather than in interlingua vocabulary. It is fairly straightforward to adapt our system to produce the meaning representation with the terminal vocabularies specified by an interlingua. However, we have made a deliberate decision to leave the Korean vocabularies in the representation largely (1) to retain the system efficiency for mapping parsing output into meaning representation, and (2) for unified execution of automation algorithms for both Korean-to-English and English-to-Korean translation. And we would like to point out that this minor compromise in meaning representation still ensures the major benefit of interlingua approach to machine translation, namely, $2 \times N$ sets of grammar rules for N language pairs, as opposed to $2^N$. 
Figure 4. Semantic Frame for the input sentence “URI Ga EoRyeoUn MunJe Reul PulEox Da.”

The semantic frame captures the core predicate-argument structure of the input sentence in a hierarchical manner, [9,10] (i.e., the internal argument, typically object, is embedded under the verb, and the external argument, typically subject, is at the same hierarchy as the main predicate, i.e., verb phrase in syntactic terms). The predicate and the arguments along with their representation categories are bold-faced in Figure 4. With the semantic frame as input, the generation system generates the English translation using the grammar rules in (1), and the Korean paraphrase using the grammar rules in (2).

The semantic frame captures the core predicate-argument structure of the input sentence in a hierarchical manner, [9,10] (i.e., the internal argument, typically object, is embedded under the verb, and the external argument, typically subject, is at the same hierarchy as the main predicate, i.e., verb phrase in syntactic terms). The predicate and the arguments along with their representation categories are bold-faced in Figure 4. With the semantic frame as input, the generation system generates the English translation using the grammar rules in (1), and the Korean paraphrase using the grammar rules in (2).

(1b) and (2b) state that the topic category for the object follows the verb predicate in English, whereas it precedes the verb predicate in Korean.

The predicate-argument structure also provides a means for word sense disambiguation, [13,15]. The verb pul_v is at least two-way ambiguous between solve and untie. Word sense disambiguation is performed by applying the rules, as in (3).

(3a) states that if the verb pul_v occurs with an object of type problem, it is disambiguated as pul+solve_v. (3b) states that the verb occurring with an object of type thread is disambiguated as pul+untie_v. The disambiguated verbs are translated into solve and untie, respectively, in the Korean-to-English translation lexicon.

1.2 Knowledge-Based Automated Acquisition of Grammars

To overcome the knowledge bottleneck for robust translation and efficient system porting in an interlingua-based system [7], we have developed a technique for automated acquisition of grammar rules which leads to a simultaneous acquisition of rules for (i) the parser, (ii) the mapper between the parser and the semantic frame, and (iii) the generator.

The technique utilizes a list of words and their corresponding parts-of-speech in the corpus as the knowledge source, presupposes a set of knowledge-based rules to be derived from a word and its part-of-speech pair, and gets executed according to the procedure given in Figure 5. The rationale behind the technique is that (i) given a word and its part-of-speech, most of the syntactic rules associated with the word can be automatically derived according to the projection principle (the syntactic
The rules for the parser for the verb *tell* in English are given below, to illustrate the dependency of the rules acquired to the specific implementation of the grammar of the language for parsing:

```
 vp_tell
  vtell [adverb_phrase] dir_object [v_pp]
  vtell [adverb_phrase] indir_object dir_object
  vtell [adverb_phrase] dir_object v_to_pp [v_pp]
  vtell [adverb_phrase] dir_object that_clause
  vtell [and_verb] [or_verb] [adverb_phrase] dir_object whClause
```

For quality evaluation, we have adopted a 5-point scale evaluation score, defined as follows. **Score 4:** Translation is both accurate and natural. **Score 3:** Translation is accurate with minor grammatical errors which do not affect the intended meaning of the input, e.g. morphological errors such as "swam vs. swimmed." **Score 2:** Translation is partially accurate, and sufficient for content understanding. Most errors are due to inaccurate word choice, inaccurate word order, and partial translation. **Score 1:** Translation is word-for-word, and partial content understanding is
possible. **Score 0:** There is no translation output, or no content understanding is possible.

We have performed the quality evaluation on 410 clauses from the training data, and 80 clauses from the test data. We have conducted the evaluation in 3 phases. **Eval 1:** Baseline evaluation after grammar and lexicon acquisition. **Eval 2:** Evaluation after augmenting word sense disambiguation rules. **Eval 3:** Evaluation after augmenting word sense disambiguation rules and accurate word order generation rules. The purpose of the 3-phase evaluation was to examine the contribution of parsing, word sense disambiguation and accurate word order generation to the overall translation quality. Once the score had been assigned to each clause, the translation score was obtained by the formula: \((\text{Sum of the scores for each clause} \times 25) / \text{Number of clauses evaluated}\).

Evaluation results are shown in Table 2 and Table 3 in terms of parsing coverage (P) and the translation score (T).

**Table 2. Translation Quality Evaluation on Training Data**

<table>
<thead>
<tr>
<th></th>
<th>Eval 1</th>
<th>Eval 2</th>
<th>Eval 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>92</td>
<td>58</td>
<td>94</td>
</tr>
<tr>
<td>T</td>
<td>79</td>
<td>55</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>69</td>
<td>69</td>
<td>74</td>
</tr>
</tbody>
</table>

**Table 3. Translation Quality Evaluation on Test Data**

<table>
<thead>
<tr>
<th></th>
<th>Eval 1</th>
<th>Eval 2</th>
<th>Eval 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>79</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>T</td>
<td>74</td>
<td>69</td>
<td>74</td>
</tr>
</tbody>
</table>

For both training and test data, the baseline translation quality score is over 50, sufficient for content understanding of the documents. Word sense disambiguation (Eval 1 vs. Eval 2) increases the translation score by about 10%, indicating that effective word sense disambiguation has a great potential for improving the translation quality.

We would like to point out that the evaluations reported in this paper are performed on clauses rather than sentences (which often consist of more than one clause). In a very recent evaluation, we have found out that evaluations on sentences decrease the overall translation score about by 15%. Nevertheless, the translation quality is still good enough for content understanding with some effort. The primary cause for the lower translation scores when the evaluation unit is a sentence as opposed to a clause is due to either an incorrect clause boundary identification, or some information (e.g., missing arguments in embedded clauses) which cannot be easily recovered after a sentence is fragmented into clauses. This has led to the ability to handle complex sentences as the primary research issue, and we are working out the solution of utilizing syntactically annotated corpus for both grammar and probability acquisition, as discussed in Section 2.3.

4. SUMMARY AND ONGOING WORK

We have described the key features of the CCLINC interlingua-based Korean-to-English translation system which is capable of translating a large quantity of Korean newspaper articles on missiles and chemical biological warfare in real time. Translation quality evaluations on the training and test data indicate that the current system produces translation sufficient for content understanding in the training domains. The key research issues identified from the evaluations include (i) parsing complex sentences, (ii) automated acquisition of word sense disambiguation rules from the training corpus, and (iii) development of discourse module to identify the referents of missing arguments. Our solution to the key technical challenges crucially draws upon the utilization of annotated corpora: For complex sentence parsing, we acquire both rules and rule production probabilities from syntactically annotated corpus. For automated word sense disambiguation, we utilize a sense-tagged corpus to identify various senses of a word, and obtain probabilities for word senses in various contexts. For discourse understanding, we are developing an algorithm for our 2-way speech translation work, [12], and plan to expand the module for document translations.

5. ACKNOWLEDGMENTS

We would like to acknowledge Dr. Jun-Tae Yoon, who provided us with a high-quality robust Korean morphological analyzer called **morany** during his stay at the Institute for Research in Cognitive Science, University of Pennsylvania as a postdoctoral fellow. **Morany** has served as a pre-processor of the understanding module in the CCLINC Korean-to-English translation system.

6. REFERENCES


