LINGSTAT: AN INTERACTIVE, MACHINE-AIDED TRANSLATION SYSTEM*

Jonathan Yamron, James Baker, Paul Bamberg, Haakon Chevalier, Taiko Dietzel, John Elder, Frank Kampmann, Mark Mandel, Linda Manganaro, Todd Margolis, and Elizabeth Steele

Dragon Systems, Inc., 320 Nevada Street, Newton, MA 02160

ABSTRACT

In this paper we present the first implementation of LINGSTAT, an interactive machine translation system designed to increase the productivity of a user, with little knowledge of the source language, in translating or extracting information from foreign language documents. In its final form, LINGSTAT will make use of statistical information gathered from parallel and single-language corpora, and linguistic information at all levels (lexical, syntactic, and semantic).

1. INTRODUCTION

The DARPA initiative in machine translation supports three very different avenues of research, including CANDIDE's fully automatic system [1,2], the interactive, knowledge-based system of the PANGLOSS group [3–6], and LINGSTAT, also an interactive system. LINGSTAT, as its name implies, incorporates both linguistic and statistical knowledge representations. It is intended for users who are native speakers of the target language, and is designed to be useful to those with little knowledge of the source (by providing access to foreign language documents), as well as those with a greater knowledge of the source (by improving productivity in translation). Although a future implementation will suggest translations of phrases and sentences, high quality automatic translation is not a goal; LINGSTAT's purpose is to relieve users of the most tedious and difficult translation tasks, but may well leave problems that the user is better suited to solve.

Initial efforts have been focused on the translation of Japanese to English in the domain of mergers and acquisitions, and a first version of a translator's workstation has been assembled. Work has also begun on a Spanish version of the system. As resources become available, particularly parallel corpora, the Spanish system will be further developed and work will be extended to include other European languages. This paper describes the Japanese system.

Japanese poses special challenges in translation that are not seen in European languages. The most striking are that Japanese text is not divided into words, and that the number of writing symbols is very large. These symbols can be divided into at least four sets: kanji, hiragana, katakana, and, occasionally, the Latin alphabet. The general-use kanji number about 2000. They are not phonetic symbols (most have several pronunciations, depending on context), but carry meaning and often appear two or three to a word. Hiragana and katakana, on the other hand, are phonetic alphabets; hiragana is usually used for important function words in Japanese grammar (sentence particles, auxiliary verbs) and to indicate inflection of verbs, adjectives, and nouns, while katakana is used almost exclusively for borrowed foreign words.

Another difficulty of Japanese is that it lacks many grammatical features taken for granted in English, such as plurals, articles, routine use of pronouns, and a future tense. Conversely, there are many Japanese concepts that have no analog in English, including the many levels of politeness, the notion of a sentence topic distinct from its subject, and exclusive *vs.* non-exclusive listings. In addition, Japanese word order and sentence structure are very different from English.

This paper is organized as follows. Section 2 lists the dictionaries and text resources used in assembling LINGSTAT. Section 3 presents an outline of the system components, some of which are described in greater detail in section 4. Section 5 describes the results of the DARPA July 1992 evaluation of the Japanese system, as well some informal results on the Spanish system. Section 6 discusses some improvements planned for future versions of the workstation.

2. RESOURCES

LINGSTAT currently makes use of a number of dictionaries and text sources of Japanese. As yet, there is no high-quality source of parallel Japanese-English text.

Dictionaries

- EDR Dictionary
  Approximately 400,000 words defined in both En-

*This work was sponsored by the Defense Advanced Research Projects Agency under contract number J-FBI-91-239.

glish and Japanese (about 200,000 distinct definitions)

- Japanese-English CD-ROM Dictionary
  Pronunciations and glosses for approximately 50,000 Japanese words
- ICOT morphological dictionary
  Pronunciations and parts of speech for approximately 150,000 Japanese words

Text

- TIPSTER articles
  Japanese newspaper articles on joint ventures
- Technical abstracts
  10,000 scientific abstracts in Japanese, with English summaries or low-quality translations
- Asahi Sinbun CD-ROM
  Seven years of Japanese newspaper articles, all subjects

3. OVERVIEW OF SYSTEM ARCHITECTURE

An initial implementation of the interactive translation system for Japanese has been completed, running under MS-DOS on PC (486) hardware. In its current form, lexical and syntactic analyses are done in a preprocessing step (initiated by the user) that produces an annotated source document and a document-specific dictionary, which are then presented to the user in a customized word-processing environment.

The pre-processing step consists of a number of subtasks, including:

1. Breaking the Japanese character stream into words using a maximum-likelihood tokenizer in conjunction with a morphological analyzer (de-inflector) that recognizes all inflected forms of Japanese verbs, adjectives, and nouns
2. Attaching lexical information to the identified words, including inflection codes and roots (for inflected forms), pronunciation, English glosses (some automatically generated from parallel text), and English definitions
3. Finding "best guess" transliterations of katakana words using dynamic-programming techniques
4. Translating numbers with following counters (eliminating a large source of user errors arising from the unusual numbering conventions in Japanese)
5. Using a finite-state parser to identify modifying phrases
6. Creating the annotated document and document-specific dictionary

The user's word-processing environment consists normally of two windows, one containing the original Japanese broken into words and annotated with pronunciation and "best guess" glosses, the other for entry of the English translation. Information extracted during pre-processing but not available in the annotated document (longer definitions, inflection information, *etc.*) can be accessed instantly from the document-specific dictionary using the keyboard or mouse, and is presented in a pop-up window. The interface also allows easy access to browsing resources such as on-line dictionaries and proper name lists.

4. IMPLEMENTATION DETAILS

Tokenization. Tokenization is done using a maximum-likelihood algorithm that finds the "best" way to break up a given sentence into words. Conceptually, the idea is to find all ways to tokenize a sentence, score each tokenization, then choose the one with the best score. The tokenizer uses a master list of Japanese words with unigram frequencies.

The score of a tokenization is defined to be the sum of the scores assigned to the words it contains, and the score of a word is taken to be proportional to the log of its unigram probability. Any character sequence not in the master list is considered infinitely bad, although to guarantee that a tokenization is always found, an exception is made for single character tokens not in the master list, which are assigned a very low, but finite, score. The tokenizer also assigns a moderate score to unfamiliar strings of ASCII or katakana, as well as to numbers.

The search for the best tokenization is done using a simple dynamic programming algorithm. Let $score(w)$ and $length(w)$ denote the score and length of the character sequence w. For a sentence of N characters numbered from 0 to $N-1$, let $best[i]$ denote the score of the best tokenization of the character sequence from 0 to $i-1$, and initialize $best[0] = 0$, $best[i] = -\infty$ for $1 < i < N$. The best tokenization score for the sentence is then given by $best[N]$ after:

FOR $i = 0$ *to* $N-1$ *DO*
  FOR all sequences w *that start at position* i *DO*
    IF $best[i] + score(w) > best[i + length(w)]$
    THEN $best[i + length(w)] = best[i] + score(w)$

Note that when two tokenizations both have a word ending at a given position, only the higher scoring solution up to that position is used in subsequent calculations.

Currently the most serious tokenization errors are caused by kanji proper nouns in the incoming document. Unlike European languages, there is no lexical cue (such as capitalization) to identify such nouns, and since most kanji can appear as words in isolation, the tokenizer will always find some way to break up a multi-kanji name into legal, but probably not sensible, pieces.

De-inflection. In order to keep the master list relatively small, only root forms of words that inflect have an entry. To recognize inflected forms, the tokenizer calls a de-inflector whenever it fails to find a candidate token in the master list.

In Japanese there are three classes of words that inflect: verbs (no person or number, but negatives and many tenses), adjectives (no cases or plurals, but negatives, adverbial, and tense), and *nani*-nouns (adjectival and adverbial). De-inflection is typically a multi-step process, as in

tabetakunakatta (didn't want to eat)
→ *tabetakunai* (doesn't want to eat)
→ *tabetai* (wants to eat)
→ *taberu* (eats).

It may also happen that a particular form can de-inflect along multiple paths to different roots.

The engine of the LINGSTAT de-inflection module is language-independent (to the extent that words inflect by transformation of their endings), driven by a language-specific de-inflection table. It handles multi-step and multi-path de-inflections, and for a given candidate will return all possible root forms to the tokenizer, along with the probability of the particular inflection for incorporation into the word score. The de-inflector also returns information about the de-inflection path for use by the annotation module. De-inflection tables have been developed for Japanese, Spanish, and English.

Annotation. The annotation module attaches pronunciations, English glosses, English definitions, and inflection information to each word identified by the tokenizer.

Pronunciation information might seem superfluous but is often of value to a Japanese translator. One of the consequences of the difficulty of written Japanese is that most students of the language can speak much better than they can read (recall that the pronunciation of a kanji cannot be deduced from its shape). The verbal cue that LINGSTAT provides through the pronunciation may therefore be enough to allow a user to identify an otherwise unfamiliar kanji word. In any case, having the pronunciation allows the user access to supplementary paper dictionaries ordered by pronunciation, which are much faster to use than radical-and-stroke dictionaries ordered by character shape information.

The glosses used by LINGSTAT come from three sources: hand entry, the Japanese-English CD-ROM dictionary, and automatic extraction from the definitions in the EDR dictionary. There are two methods of automatic extraction:

- Pull the gloss out of the definition—for example, *A type of financial transaction named leveraged buyout* becomes *leveraged buyout.*
- Use the English and Japanese definitions in the EDR dictionary as sentenced-aligned parallel text and apply CANDIDE's word alignment algorithm (Model 1) [1] to determine which English words correspond to each Japanese word.

The first method is moderately successful because many of the definitions adhere to a particular style. The second method gives good glosses for those Japanese words that occur frequently in the text of the definitions.

Katakana Transliteration. Words are borrowed so frequently from other languages, particularly English, that their transliterations into katakana rarely appear in even the largest dictionaries. The best way to determine their meaning, therefore, is to transliterate them back into English. This is made difficult by the fact that the transformation to katakana is not invertible: for example, English *l* and *r* both map to the Japanese *r*, *r* following a vowel is sometimes dropped, and vowels are inserted into consonant clusters.

The LINGSTAT katakana transliterator attempts to guess what English words might have given rise to an unfamiliar katakana word. It converts the katakana pronunciation into a representation intermediate between Japanese and English, then compares this to a list of 80,000 English words in the same representation. A dynamic programming algorithm is used to identify the English words that most closely match the katakana. These words are then attached to the katakana token in the annotation step.

This procedure fails for non-English foreign words, and for most proper names (since they rarely appear in the master English list).

Number Translation. In traditional Japanese, numbers up to 10^4 are formed by using the kanji digits in

conjunction with the kanji symbols for the various powers of ten up to 1000, *e.g.*, 6542 would be written

$$(6)(1000)(5)(100)(4)(10)(2),$$

with each number in parentheses replaced by the appropriate kanji symbol. Notice that the powers of ten are explicitly represented, rather than being implied by position.

There are special kanji for the large numbers 10^4, 10^8, *etc.* These may be preceded by expressions like that above to form very large numbers, such as

$$\begin{aligned}(2)(10^8)(5)(1000)(5)(100)(10^4) &= 2 \times 10^8 + 5500 \times 10^4 \\ &= 255,000,000.\end{aligned}$$

Modern Japanese often mixes the traditional Japanese representation with the "place-holding" representation used in English. Arabic numerals are freely mixed with kanji symbols in both formats. To ease the burden on the translator LINGSTAT has a function that recognizes numbers in all their styles, including following counters, and translates them into conventional English notation. These translations are then attached to the number token in the annotation step. Comparison of manual and LINGSTAT-aided translations has demonstrated that this feature eliminates a large source of critical errors, particularly in the evaluation domain, which frequently references large monetary transactions.

Finite-state parser. As a first pass at helping the user with Japanese sentence structure, LINGSTAT incorporates a simple finite-state parser designed to identify modifying phrases in Japanese sentences. An interface function has also been added to display this information in a structured way. At this stage, the quality of the parse is only fair. This function has not yet been tested for its effect on translation speed.

5. RESULTS

The system as described here (without the finite-state parser) was evaluated by DARPA in July 1992. The performance of two Level 2 translators was measured on a test set of 18 Japanese documents, each translator translating 9 with the aid of the system and 9 by hand. In general, the quality of translation with and without the system was found to be comparable, but the system provided a speedup of approximately 30%.

Since the tested system provided no help with the analysis of the Japanese sentences, this savings was achieved by drastically reducing the time spent doing tokenization and lookup. It might appear surprising that so much time could be saved from these activities alone, but the many unusual features of Japanese described above conspire to produce a large overhead in this phase of translation compared to other languages. This result is also consistent with an analysis of how the translators allocated their time: without the system, their principal effort involved dictionary lookup, but with the system most of their time was spent analyzing sentence structure.

Productivity tests have also been conducted on the rudimentary Spanish version of the workstation. This system incorporates a Spanish de-inflector, provides word for word translation to English, and has fast access to an on-line dictionary. On a scaled down version of the DARPA test (6 documents instead of 18, including 3 by hand and 3 with the aid of the system), a fluent speaker of Italian (a language very similar to Spanish) showed no productivity gain. At the other extreme, a user with no Spanish knowledge and no recent training in any European language was about 50% faster using the system's on-line tools than with a paper dictionary.

6. CURRENT AND FUTURE WORK

There are currently two programs underway to improve the translation system. The first is an effort to expand the Japanese and Spanish dictionaries, which requires not only adding words, but also glosses, pronunciations (for Japanese), and multi-word objects. Part of this task involves updating the Japanese and Spanish word frequency statistics, which will improve the performance of the tokenizer in Japanese and the de-inflector in both languages. Part of speech information is also being added, in anticipation of the use of grammatical tools.

The second program is the development of a probabilistic grammar to parse the source and provide grammatical information to the user. This will supplement or replace the current rule-based finite-state parser currently implemented in the system. In the current phase, we have chosen a lexicalized context-free grammar, which has the property that the probability of choosing a particular production rule in the grammar is dependent on headwords associated with each non-terminal symbol. Lexicalization is a useful tool for resolving attachment questions and in sense disambiguation. This grammar will be trained using the inside-outside algorithm [7] on Japanese and Spanish newspaper articles.

One use of the grammar will be to provide more accurate glossing of the source by making use of co-occurrence statistics among the phrase headwords. This requires developing an English word list with frequency and part

of speech information, as well as constructing an English inflector-deinflector. These tools, along with an English grammar, will enable the system to construct candidate translations of Japanese phrases and simple Spanish sentences.

A longer term goal of the syntactic analysis (particularly when more languages are incorporated) is to generate a probability distribution in a space of data structures in which the order of representation of the component grammatical elements is language neutral. This can regarded as a kind of syntactic interlingua. There will also be a deeper semantic analysis of the source which will be less dependent on the syntactic analysis, and will use a probabilistic model to fill in the components of a case-frame semantic interlingua. These kinds of structures will allow faster inclusion of new languages and domains.

References

1. P.F. Brown, S.A. DellaPietra, V.J. DellaPietra, and R.L. Mercer, "The Mathematics of Machine Translation: Parameter Estimation," submitted to Computational Linguistics, 1991.
2. P.F. Brown, S.A. DellaPietra, V.J. DellaPietra, J. Lafferty, and R.L. Mercer, "Analysis, Statistical Transfer, and Synthesis in Machine Translation," submitted to TMI-92, Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, 1992.
3. D. Farwell and Y. Wilkes, "ULTRA: A Multi-lingual Machine Translator," Proceedings of the Third MT Summit, pp. 19–24, 1991.
4. E. Hovy and S. Nirenburg, "Approximating an Interlingua in a Principled Way," Proceedings of the Speech and Natural Language Workshop, pp. 261–266, 1992.
5. K. Knight, "Building a Large Ontology for Machine Translation," Proceedings of the ARPA Workshop on Human Language Technology, 1993.
6. R. Frederking, D. Grannes, P. Cousseau, and S. Nirenburg, "A MAT Tool and Its Effectiveness," Proceedings of the ARPA Workshop on Human Language Technology, 1993.
7. J.K. Baker, "Trainable Grammars for Speech Recognition," Speech Communication Papers for the 97th Meeting of the Acoustical Society of America (D.H. Klatt and J.J. Wolf, eds.), pp. 547–550, 1979.