System Demonstration

TOWARDS A MULTILINGUAL ANALYST'S WORKSTATION: TEMPLE[1]

Rémi Zajac
Computing Research Laboratory

1 System builders and contacts

Dr. Rémi Zajac
Computing Research Laboratory, New-Mexico State University
Box 30001, Dept. 3CRL, Las Cruces NM 88003, USA
Tel: +1-505-646-5782, Fax: +1-505-646-6218, E-mail: zajac@crl.nmsu.edu, http://crl.nmsu.edu

2 System Category

The Temple system is a research prototype developed as a proof-of-concept for Glossary-Based Machine-Translation system, as a research vehicle for Machine Translation, and as a demonstration platform for a variety of integrated NLP tool suites. It has been installed at the DoD (sponsor of the Temple project).

3 System Characteristics

Translation speed	From 46 sec. to 67 sec. average for a page (250 words) depending on the language[a]
Domains covered	News article
Input format	Plain text
Output quality	Low, but deemed sufficient for information analysis purposes

a. *Except for the Russian-English, still under development, which takes over 3 minutes.*

4 Resources

	Dictionary	Glossary
Arabic-English	72,000 stems	11,000 phrases
Japanese-English	40,000 stems	5,000 phrases
Russian-English	80,000 stems	1,600 phrases
Spanish-English	60,000 stems	20,000 phrases

5 Hardware and Software

Platform	Sun
Operating systems	SunOS, Solaris
Programming languages	C, Lisp, Prolog, Tcl/Tk, Perl

6 Functionality Description

The Temple system provides the end-user (analyst) with the following functionalities:

- Document management.
- Document browser/editor.
- Lexical database browser/editor.
- MT functionalities from Arabic, Japanese, Russian and Spanish to English.

1. Funded by the DoD, Maryland Procurement Office, Fort George G. Meade, MD under grant 94-R-3075/A0001.

The system architecture is open and a variety of other tools developed in different projects have been integrated for various purposes, including: information retrieval and extraction tools, translation memory, on-line versions of paper dictionaries, multilingual key word in context (KWIC) and concordance packages, etc.

7 System Internals

7.1 Overview

The intended users of the Temple Translator's Workstation are translators and analysts who routinely browse, analyze and translate large numbers of documents. Each document is relatively small (typically, newspaper articles). The translator's workstation is used concurrently with other information extraction and retrieval tools, e.g. Inquery [Callan et al. 92]. The Temple Translator's Workstation is integrated in the Tipster Document Management [Grishman 95] system developed at CRL, which provides text management functionalities and allows a smooth integration of the Temple Translator's Workstation in the user's working environment.

The tools of the Temple workstation include a Unicode-based multilingual text editor (Fig. 2) with basic editing capabilities for most of the languages supported by the Unicode standard [Unicode 91], an on-line dictionary editor for modifying and enhancing the bilingual dictionaries in the system, and an on-line glossary editor (Fig. 3) for developing bilingual glossaries. The ultimate goal is that every tool in the workstation be Unicode-based. So far, however, only the editor and the Arabic-English machine translation components (including the dictionaries and the analyzers) are based on Unicode.

Temple provides an architecture for integrating Machine Translation (MT) components; the Glossary-Based Machine Translation (GBMT) system on which it relies was first developed at Carnegie Mellon University as a part of the Pangloss project [Frederking et al., 93, Nirenburg et al., 93]. In that effort, a sizeable Spanish-English glossary-based MT system was implemented. The Temple project has built upon this experience and extended the GBMT approach to other languages: Japanese, Arabic, and Russian. This experience with other languages has provided significant insights for development of a new versatile GBMT engine and for the use of off-the-shelf components for building a complete MT system.

7.2 The Multilingual MT Architecture

The machine translation architecture is based on the Tipster Document Architecture [Grishman 95], which allows flexible integration of NLP tools to build applications through a unifying mechanism: tools communicate through *annotations* on the document. Each annotation is associated with a span of text and contains information represented as objects. Thus, not only machine translation systems but also other NLP tools like taggers, spell-checkers, etc., can work concurrently on the same document. To facilitate tool integration, the Temple architecture defines a standard way of representing linguistic information, and also defines a "neutral" common linguistic structure for the purpose of machine translation that is shared by all tools.

The machine translation strategy (Fig. 1) is kept simple. The first phase of the machine translation process is a morphological analysis of the text,[1] followed by a bilingual dictionary look-up.[2] The next phase is a glossary-based transfer which uses both a bilingual glossary and the result of the dictionary look-up to produce a phrase-by-phrase translation, falling back on a word-by-word translation as needed. The result is

1. The Spanish part-of-speech tagger developed at CRL [Farewell et al., 94] and the Juman morphological analyzer developed at Kyoto University [Matsumoto et al., 93] for Japanese are used. Morphological analyzers for Arabic and Russian were developed by CRL in the Temple project.
2. Most of Temple's bilingual dictionaries are extracted from machine-readable versions of paper dictionaries [Stein et al., 93].

sent to the English morphological generator [Penman 88]. The final step is the creation of a new document containing the translation. Alternative translations are stored as annotations on the text and can be displayed using the annotation viewer. The human translator can then use the editor (Fig. 2) for producing the final translation.

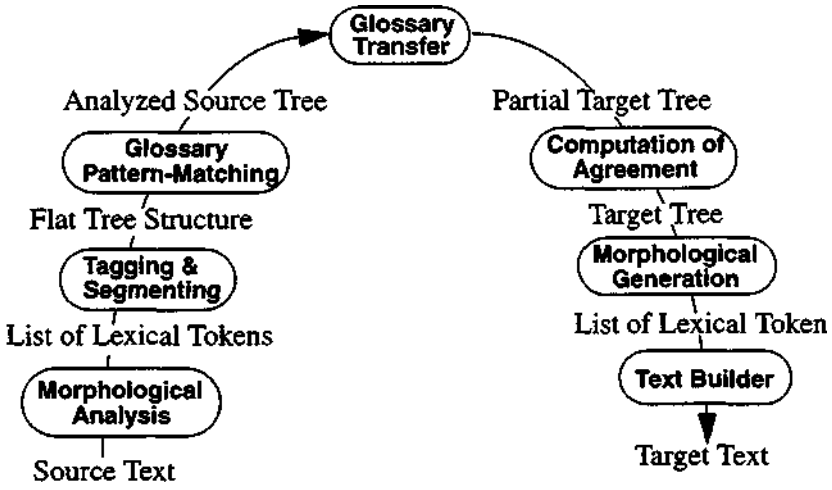

Figure 1: Process of Glossary-Based Machine Translation.

Thus, the linguistic components of a machine translation module for a given language pair are few and simple: a morphological analyzer, a bilingual dictionary, a bilingual glossary and a morphological generator.

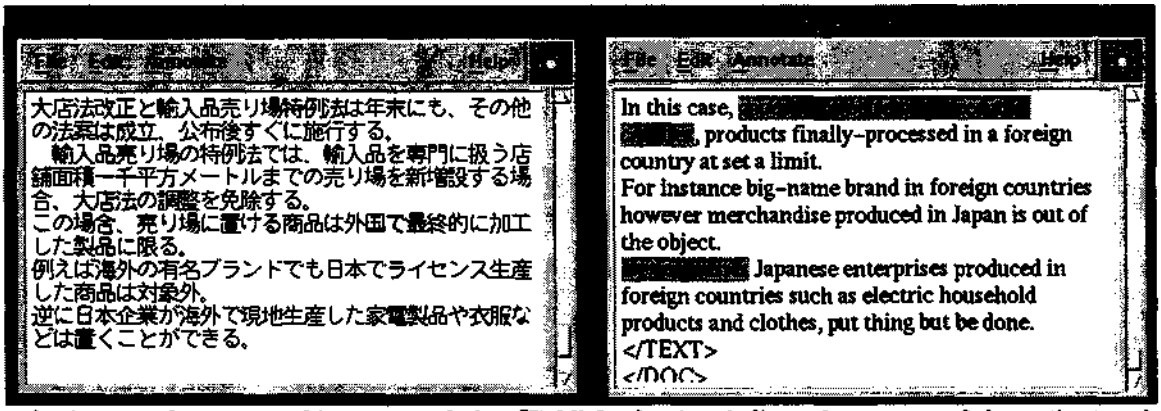

Figure 2: A source document and its raw translation. Highlighted strings indicate the presence of alternative translations.

7.3 Glossary-Based Machine Translation

The structure of a glossary entry is simple: a source phrase pattern plus a list of corresponding target patterns. A glossary entry can easily be added to the system by a user who has minimal linguistic training. Fig. 3 shows a sample screen from the glossary editor. A typical glossary entry looks like this Spanish-English pair:

```
dar<:1> <number:2> pesos de recompensa
    give<:1> a <number:2> pesos of reward
```

Applying this pattern, the source sentence "Di diez mil pesos de recompensa" is translated as "I gave a ten thousand pesos of reward".

Patterns in the glossary are matched to a sentence word by word. A word can be specified as optional in the pattern. If a sub-string is matched by several glossary entries, only the longest match is considered. Patterns are indexed and sorted according to their length. The pattern matcher records transfer information for successful matches in a tree structure, proceeding to the next unmatched word, and never backtracking. Thus, the pattern-matching algorithm is very fast and its typical behavior is linear in the number of input words.[1]

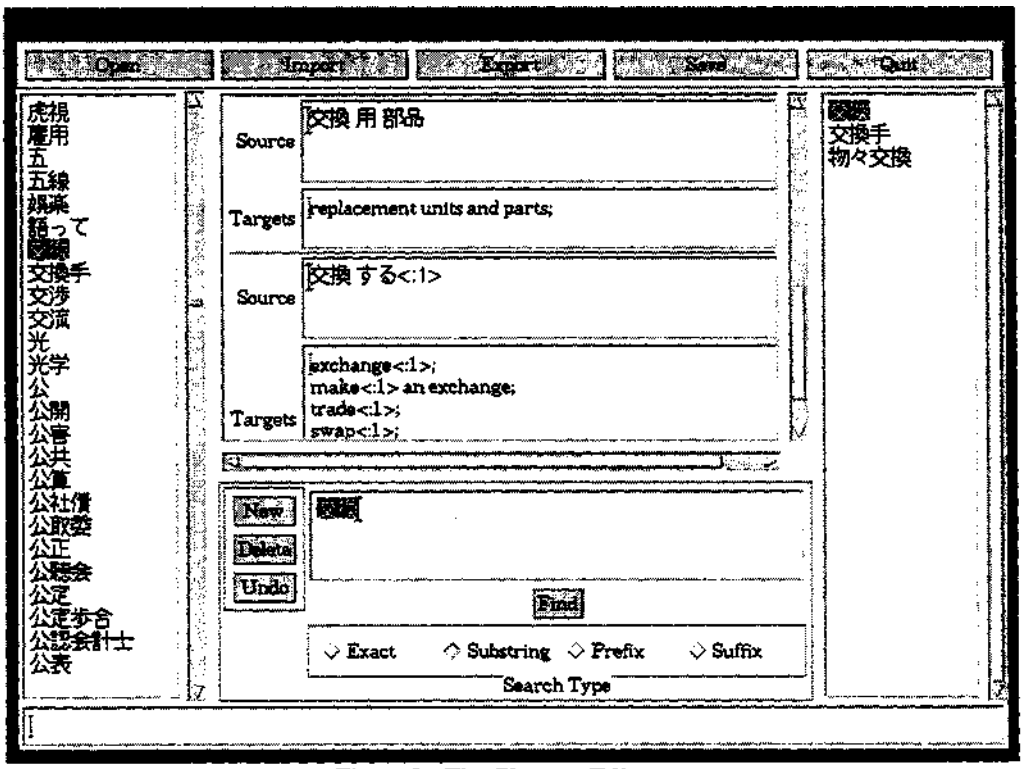

Figure 3: The Glossary Editor.

A word pattern in the glossary is typically a dictionary form followed by category specification, transfer specification and agreement specification (in between angle brackets). It is possible to specify a fully inflected form. The pattern-matcher always tries to match the dictionary form produced by the morphological analyzer before trying to match the word to the full form. To resolve ambiguities, category specification is used to select one of the analyses produced by the morphological analyzer, since category distinction usually reflects differences in semantics (for example, the difference between `hold<noun>` and `hold<verb>`).

The word pattern can also be specified by category only, like `<number:2>`, which matches any number. Function word patterns are usually specified by using the category. The category can be refined using a list of feature values that must then be a subset of the features of the lexical token

1. A previous in-memory implementation used a trie for indexing the whole glossary. The current implementation uses a persistent store and indexes only on the first word of the pattern with better overall performance, essentially due to lower consumption of memory.

8 Conclusion

The goals of this 2 years project were to:

- Develop a multilingual document architecture for natural language processing, and more specifically for multilingual machine translation;
- To develop a machine translation model supporting fast development of a machine translation systems for new languages with minimal resources and with an emphasis on reuse of NLP components and resources;
- Demonstrate the approach on several languages, Arabic, Japanese, Russian and Spanish to English.

A Web version of the original Temple interface has been developed and can be accessed through the CRL Web site (http://crl.nmsu.edu). For demonstration purposes, a version for translating foreign Web pages in English is also accessible.

A new project, the Corelli project, has been launched to develop a more generic and more robust version of the GBMT architecture. The project goals also include the development of a translation editor integrated in a commercial document processing package, and the development of a multilingual toolkit for glossary acquisition.

9 References

Callan J.P., Croft W.B., and Harding S.M., 1992. "The INQUERY Retrieval System," Proceedings of the *3rd International Conference on Database and Expert Systems Applications.* Springer-Verlag. pp-78-83.

Frederking, R., D. Grannes, P. Cousseau, and S. Nirenburg. 1993. "An MAT Tool and Its Effectiveness". Proceedings of the *DARPA Human Language Technology Workshop,* Princeton, NJ.

Farwell, D., Helmreich, S., Jin, W., Casper, M, Hargrave, J., Molina-Salgado, H. Weng, F. 1994. Panglyzer: Spanish Language Analysis System. In Proceedings of the *Conference of the Association of Machine Translation in the Americas,* AMTA'94. Columbia, MD.

Grishman, Ralph, editor. 1995. "Tipster Phase II Architecture Design Document, Version 1.52", New-York University. (http://cs.nyu.edu/tipster)

"The Penman Primer, User Guide, and Reference Manual". 1988. Unpublished USC/ISI documentation.

Matsumoto, Yuji, Sadao Kurohashi, Takeji Utsuro, Yutaka Myougi, Makoto Nagao. 1993. "Japanese Morphological Analysis System, JUMAN". Kyoto University, Nara Science and Technology Graduate School University. (in Japanese).

Nirenburg, S., P. Shell, A. Cohen, P. Cousseau, D. Grammes, C. McNeilly. 1993. "Multi-purpose Development and Operations Environments for Natural Language Applications". Proc. of the *3rd Conference on Applied Natural Language Processing* (ANLP-93), Trento, Italy.

Stein, Gees C., Lin, Fang, Bruce, Rebecca, Weng, Fuliang, and Guthrie, Louise. 1993. "The Development of an Application Independent Lexicon: LexBase". *CRL Technical Report MCCS-92-247.*

The Unicode Consortium. 1991. *The Unicode Standard, Worldwide Character Encoding.* Addison-Wesley Publishing Company.