System Demonstration

THE PANGLOSS-LITE MACHINE TRANSLATION SYSTEM

Robert E. Frederking and Ralf D. Brown

Center for Machine Translation
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890 USA

System builders and contacts: Same; ref+@cs.cmu.edu, ralf+@cs.cmu.edu

System category: Research vehicle

System characteristics: 4-8 seconds per spoken sentence, multiple or unrestricted domains

Resources: See Figure 2-1.

Hardware and software: Sun Sparcstations w/SunOS; Intel PCs w/ Microsoft Windows NT or 95

Functionality description: Employed in DIPLOMAT rapid-deployment speech-to-speech MT system; bidirectional in both Spanish/English and Serbo-Croatian/English; soon Korean/English

System internals: Based on multi-engine MT, EBMT, glossaries, and statistical language modelling

1. Pangloss-Lite Overview

The Pangloss-Lite (PanLite) machine translation system is a standalone C++ re-implementation of several major components from the Pangloss machine translation system [Nirenburg et al. 95][1]. It incorporates the Pangloss Example-Based MT (EBMT) [Brown 96a] and Transfer-Based MT engines, and its statistical language modeller [Brown and Frederking 95], as well as a newly-implemented morphological analyzer, within the multi-engine MT architecture [Frederking and Nirenburg 94] developed during the course of the project.

Due to improved design and the C++ implementation, PanLite runs very quickly. For example, the EBMT engine formerly required several minutes to translate a typical newswire sentence; it now requires about 15 seconds (and this with a much larger corpus). More details on performance are presented in section 2 below.

To allow its use in the widest variety of applications, PanLite has been designed to translate strings provided either on the standard input or via network sockets, and to produce as output either the best

[1]Pangloss was a joint project between three sites: the Computing Research Laboratory of New Mexico State University, the Information Sciences Institute of the University of Southern California, and the Center for Machine Translation of Carnegie Mellon University. It was funded by the U.S. Department of Defense.

composite string or the full chart of scored translated segments. The latter is necessary, for example, when the output will be supplied to an external graphical user interface (GUI) for post-editing.

PanLite has already been included as the MT component of the prototype DIPLOMAT rapid-deployment speech-to-speech translation system (see section 3, below). A potential future application of PanLite is as a World Wide Web translation server.

1.1. Multi-Engine Machine Translation

The overall organization of PanLite is shown in Figure 1-1. PanLite employs a multi-engine MT architecture [Frederking and Nirenburg 94]: several MT engines, each employing a different MT technology, are applied in parallel to each input text. Each engine attempts to translate the entire input text, segmenting each sentence in whatever manner is most appropriate for its technology, and putting the resulting output segments into a shared chart data structure after giving each segment a score indicating the engine's internal assessment of the quality of the output segment. The output segments are indexed in the chart based on the positions of the corresponding input segments. Since the scores produced by the engines are not very reliable, we use statistical language modelling techniques adapted from speech recognition research to select the best overall set of outputs [Brown and Frederking 95].

1. Text input via standard input or sockets
2. Morphological analysis
3. Translation: results of morphological analysis passed to each MT engine; scored outputs placed into chart
4. Language modeller selects ''best'' edges, and adds results to chart
5. Output: either text composed of ''best'' edges or entire chart

Figure 1-1: Structure of PanLite

In PanLite, the translation engines used are:

- **EBMT:** EBMT [Brown 96a] uses a sentence-aligned corpus to produce translations. When such a corpus is available, fairly high-quality MT for a new domain is available essentially immediately. EBMT is basically a more sophisticated version of Translation Memory, in that sub-sentential chunks of words are matched, allowing much greater coverage. Sentences that match in full are translated exactly, but sub-sentential chunks are matched with a variety of heuristics, which are reflected in the scores assigned to them. The greatly increased speed of the PanLite C++ implementation has allowed the entirety of the largest available corpora to be indexed and used for EBMT, something that had not been feasible previously.
- **Transfer-based MT:** This engine employs a very simple, very old technology: bilingual dictionaries and phrasal glossaries are used to translate pieces of source text. While this is a low-quality technique, the simplicity of the technique allows us to quickly and semi-automatically develop large databases, allowing an initial rapid-deployment of an MT system while more sophisticated KBMT engines are developed. Also, any available online bilingual dictionaries can be used immediately. Scores are currently statically assigned on a per-glossary basis, based on our overall confidence in the particular glossary. An important development in PanLite is the merging of the code implementing glossaries and EBMT, significantly simplifying further software development and maintenance.

- **Knowledge-Based MT:** Currently the PanLite system does not contain a Knowledge-Based MT (KBMT) engine, although a slot is already present to add one later. To be suitable for integration with the other engines, the KBMT system should preferably produce translations as quality-scored segments of sentences, as the Pangloss KBMT engine does, rather than only full sentences.

1.2. Morphological analysis

PanLite is designed to use morphological analysis as in its predecessor Pangloss system, to produce stem forms and feature taggings for all the words of the input, before they are passed to the different engines. KBMT requires such analysis, and EBMT and transfer-based MT can also clearly benefit from it. Our group is currently producing a C++ version of the Morphe morphological analyzer [Leavitt 94], iCelos, for use in this system. Pending its completion, its output is augmented using a file containing an indexed list of stems or roots for each source language.

1.3. Language Modelling

As mentioned above, we use statistical language modelling to combine the segments produced by different engines, a technique borrowed from speech recognition work. There, acoustic recognizers produce many hypotheses for each word, with scores that are not very accurate. Quality is improved by applying a statistical language model to such results. The model is produced by analyzing large amounts of English text to see what the most probable sequences of words are in English. The model is then used to find the set of choices that produces the sequence most likely to be an English sentence, taking into account the scores of the component words. We use a trigram model of the target language, with backoff to bigrams and unigrams. That is to say, we use the probabilities of word triples when we have these available. When the trigram probability is unavailable, we use the probabilities of word pairs or single words. Because of the extremely large number of combinations of segment hypotheses, search becomes necessary, as described in [Brown and Frederking 95].

2. PanLite System Details

Currently, versions of PanLite exist for translating unrestricted Spanish to English, Serbo-Croatian to English, English to Spanish, and English to Serbo-Croatian. The code is the same for each version, with just databases and configuration files changing. The sizes of the code and the various databases are presented in Figure 2-1. The FramepaC library [Brown 96b] provides frame-based and Lisp-like data structure capabilities. PanLite currently runs on Sun Sparcstations under SunOS and on Intel processors under Microsoft Windows NT or Windows 95, and the runtime databases are binary-compatible between platforms.

Performance figures for the EBMT system on a Sun Sparcstation LX are illustrative: a sample Spanish newswire text of 15 sentences totalling 414 words and punctuation marks can be translated in just under four minutes (see also Figure 2-2). 20 texts averaging 450 words each, drawn from the ARPA MT evaluations, can be completely processed in about three hours, including dictionary lookups and statistical modeling (that is, all processing except the glossaries).

Indexing the entire 280M Spanish-English EBMT corpus requires approximately 45 minutes on a Sparcstation LX when all files are located on local disks, and another 30 minutes to pack the index (not required, but improves speed at run time). Incremental addition of new data to the corpus proceeds at a rate of roughly six megabytes per minute.

The bilingual Spanish-English corpus consists of 726,406 sentence pairs drawn primarily from the UN Multilingual Corpus [Graff and Finch 94], with a small admixture of texts from the Pan-American Health Organization and the ARPA MT evaluations (10250 sentence pairs stem from the PAHO corpus and 552 pairs from evaluations). The Serbo-Croatian/English corpus is currently much smaller at only 34,000 pairs, drawn from online parallel texts, scanned-in bilingual newspapers, and the glossaries.

Code:
PanLite main program: 4,500 lines of code
EBMT/glossary: 12,300 lines of code
LM: 9,700 lines of code
FramepaC: 50,600 lines of code
(used by all three programs)
Total object code size: about 1200K for SunOS and 900K for Windows NT.

Data:
PanLite:
39,800-word Serbo-Croatian stem list
12,300-word English root list
41,300-word Spanish root list
EBMT:
280M Spanish-English corpus
280M English-Spanish corpus (inverse of S-E)
2.3M SerboCroatian-English corpus
2.3M English-SerboCroatian corpus (inverse of SC-E)
19,700-word English root/synonym list
56,900-word Spanish-Eng association dictionary
21,300-word Eng-SCro association dict
51,100-word SCro-Eng association dict
Glossaries:
193,000-entry Spanish-English glossary
85,000-entry SerboCroatian-English glossary
129,000-entry English-SerboCroatian glossary
(SC-E and E-SC glossaries contain an MRD)
Language Modeller:
13M Serbo-Croatian model (from about 12M text)
60M English model (from about 450M text)
41M Spanish model (from about 135M text)

Figure 2-1: Code and database sizes

Croatian-English/English-Croatian:

Sparcstation LX:	10-15 seconds
Windows NT/95 (Pentium-90):	4-8 seconds

Spanish-English/English-Spanish:

Sparcstation LX:	15-25 seconds

Figure 2-2: Times to Translate Typical Sentences

3. Rapid Deployment MT

The PanLite system is the translation component for the DIPLOMAT rapid-deployment, wearable speech-to-speech translation project. One of DIPLOMAT's goals is ''rapid-deployment'': being able to perform initial translations of a new language in a matter of days or weeks. The initial version of the DIPLOMAT bidirectional Serbo-Croatian/English prototype system, which we will be demonstrating on Toshiba laptops, was developed from scratch in less than three weeks.

The language-pair-independence of the software was been further demonstrated recently: English-to-Spanish translation was brought up on July 29, 1996 in less than seven hours. During these seven hours, a single person using a single Sun Sparcstation inverted the existing Spanish-to-English corpus, dictionary, and glossaries; created new configuration files; created a Spanish language model; and indexed the EBMT corpus, the dictionary, and the glossaries. While the initial translations are of lower quality than the Spanish-to-English translations (due primarily to the poor quality of the inverted dictionary), they can be improved incrementally with some additional effort. Of course, this exercise finessed a number of difficult issues that the full project is addressing, especially the rapid development of the knowledgebases for a completely new language. But it does demonstrate the generality of the software, and that knowledgebase development *is* the primary remaining MT challenge.

4. References

[Brown 96a] Brown, R.D. 1996. ''Example-Based Machine Translation in the Pangloss System.'' In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-96).

[Brown 96b] Brown, R.D. 1996. ''FramepaC User's Manual,'' Carnegie Mellon University Center for Machine Translation technical memorandum (in preparation). Current draft available as http://www.cs.cmu.edu/afs/cs.cmu.edu/user/ralf/pub/WWW/papers.html.

[Brown and Frederking 95] Brown, R. and Frederking, R. 1995. ''Applying Statistical English Language Modeling to Symbolic Machine Translation.'' In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-95), pp. 221-239.

[Frederking and Nirenburg 94] Frederking, R. and Nirenburg, S. 1994. ''Three Heads are Better than One.'' Proceedings of the fourth Conference on Applied Natural Language Processing, ANLP-94, Stuttgart, Germany.

[Graff and Finch 94] Graff, D. and Finch, R. 1994. ''Multilingual Text Resources at the Linguistic Data Consortium.'' In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Morgan Kaufmann.

[Leavitt 94] Leavitt, J. 1994. ''Morphe: A Morphological Rule Compiler,'' Version 2.0a, CMU-CMT-94-MEMO. Carnegie Mellon University Center for Machine Translation technical memorandum.

[Nirenburg et al. 95] Nirenburg, S., (ed.). 1995. ''The Pangloss Mark III Machine Translation System.'' Joint Technical Report, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie Mellon University), Information Sciences Institute (University of Southern California). Issued as CMU technical report CMU-CMT-95-145.

Table of Contents

List of Figures