Example-Based Machine Translation in the Pangloss System

Ralf D. Brown
Center for Machine Translation
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
ralf@cs.cmu.edu

Abstract

The Pangloss Example-Based Machine Translation engine (PanEBMT)[1] is a translation system requiring essentially no knowledge of the structure of a language, merely a large parallel corpus of example sentences and a bilingual dictionary. Input texts are segmented into sequences of words occurring in the corpus, for which translations are determined by subsentential alignment of the sentence pairs containing those sequences. These partial translations are then combined with the results of other translation engines to form the final translation produced by the Pangloss system. In an internal evaluation, PanEBMT achieved 70.2% coverage of *unrestricted* Spanish news-wire text, despite a simplistic subsentential alignment algorithm, a suboptimal dictionary, and a corpus from a different domain than the evaluation texts.

1 Introduction

Pangloss (Nirenburg et al., 1995) is a multi-engine machine translation system, in which several translation engines are run in parallel to propose translations of various portions of the input, from which the final translation is selected by a statistical language model. PanEBMT is one of the translation engines used by Pangloss.

EBMT is essentially translation-by-analogy: given a source-language passage S and a collection of aligned source/target text pairs, find the "best" match for S in the source-language half of the text collection, and accept the target-language half of that match as the translation. PanEBMT, like other example-based translation systems, uses essentially no knowledge about its source or target languages; what little knowledge it does use is optional, and is supplied in a configuration file. Its three main knowledge sources are: a sententially-aligned parallel bilingual corpus; a bilingual dictionary; and a target-language root/synonym list. The fourth (minor and optional) knowledge source is the language-specific information provided in the configuration file, which consists of a list of tokenizations equating words within classes such as weekdays, a list of words which may be elided during alignment (such as articles), and a list of words which may be inserted

[1]This work as part of the Pangloss project was supported by the U.S. Department of Defense

2 Parallel Bilingual Corpus

The corpus used by PanEBMT consists of a set of source/target sentence pairs, and is fully indexed on the source-language sentences. The corpus is not aligned at any granularity finer than the sentence pair; subsentential alignment is performed at run-time based on the sentence fragments selected and the other knowledge sources.

The corpus index lists all occurrences of every word and punctuation mark in the source-language sentences contained in the corpus. The index has been designed to permit incremental updates, allowing new sentence pairs to be added to the corpus as they become available (for example, to implement a translation memory with the system's own output). The text is tokenized prior to indexing, so that words in any of the equivalence classes defined in the EBMT configuration file (such as month names, countries, or measuring units), as well as the predefined equivalence class **<number>**, are indexed under the equivalence class rather than their own names. For each distinct token, the index contains a list of the token's occurrences, consisting of a sentence identifier and the word number within the sentence. At translation time, PanEBMT back-substitutes the appropriate target-language word into any translation which involves any tokenized words.

The bilingual corpus used for the results reported here consists of 726,406 Spanish-English sentence pairs drawn primarily from the UN Multilingual Corpus available from the Linguistic Data Consortium(Graff and Finch, 1992) (Figure 1), with a small admixture of texts from the Pan-

```
Las fuentes de esos comentarios y
  recomendaciones son las siguientes :
The sources of these comments and
  recommendations are :
El informe de la Junta de Auditores a la
  Asamblea General que incluye las
  observaciones del Director Ejecutivo
  del UNICEF sobre los comentarios y
  recomendaciones de la Junta de
  Auditores ;
The report of the Board of Auditors to
  the General Assembly which incorporates
  the observations of the Executive
  Director of UNICEF on the comments and
  recommendations of the Board of
  Auditors ;
```

Figure 1: Corpus Sentence Pairs

```
(ACADMICOS ACADEMICS ACADEMICAL
  TITLES DEGREES)
(ACAECIDO HAPPEN)
(ACAECIDOS HAPPEN)
(ACANTONADAS CANTON QUARTER TROOPS)
(ACANTONAMIENTO CANTONMENT)
(ACARREA CARRY CART HAUL TRANSPORT
  CAUSE OCCASION)
(ACARREABA CARRY CART HAUL TRANSPORT
  CAUSE OCCASION)
    ...
(ACARREARON CARRY CART HAUL TRANSPORT
  CAUSE OCCASION)
(ACARREAR TRANSPORT HAUL CART CARRY
  LUG ALONG BRING DOWN CAUSE OCCASION
  ITS TRAIN RESULT GIVE RISE)
```

Figure 2: Bilingual Dictionary Entries

American Health Organization and prior project evaluations[2], indexed as described above.

Together, the bilingual dictionary and target-language list of roots and synonyms (extracted from WordNet when translating into English) provide the necessary information to find associations between source-language and target-language words in the selected sentence pairs. These associations are used in performing subsentential alignment. A source word is considered to be associated with a target-language word whenever either the target word itself or any of the words in its root/synonym list appear in the list of possible translations for the source word given by the dictionary.

Not all words will be associated one-to-one; however, the current implementation requires that at least one such unique association be found in order to provide an anchor for the alignment process.

3 Implementation

PanEBMT is implemented in C++, using the FramepaC library (Brown, 1996) for accessing Lisp data structures stored in files or sent from the main Pangloss module via Unix pipes. PanEBMT consists of approximately 13,300 lines of code, including the code for a glossary mode which will not be described here.

PanEBMT uses a re-processed version of the bilingual dictionary used by Pangloss's dictionary translation engine (Figure 2). The re-processing consists of removing various high-frequency words and splitting all multi-word definitions into a list of single words, needed to find one-to-one associations.

[2]10250 sentence pairs stem from the PAHO corpus and 552 pairs from evaluations.

4 EBMT's Place in Pangloss

PanEBMT is merely one of the translation engines used by Pangloss; the others are transfer engines (dictionaries and glossaries) and a knowledge-based machine translation engine (Figure 3). Each of these produces a set of candidate translations for various segments of the input, which are then combined into a chart (Figure 3). The chart is passed through a statistical language model to determine the best path through the chart, which is then output as the translation of the original input sentence.

5 EBMT Operation

The EBMT engine produces translations in two phases:

1. find chunks by searching the corpus index for occurrences of consecutive words from the input text
2. perform subsentential alignment on each sentence pair found in the first phase to determine the translation of the chunk

In constrast with other work on example-based translation, such as (Maruyama and Watanabe, 1992) or early Pangloss EBMT experiments (Nirenburg et al., 1993), PanEBMT does not find an optimal partitioning of the input. Instead, it attempts to produce translations of every word sequence in the input sentence which appears in its corpus. The final selection of the "correct" cover for the input is left for the statistical language model, as is the case for all of the other translation engines in Pangloss. An advantage of this approach is that it avoids discarding possible chunks merely because they are not part of the "optimal" cover for the input, instead selecting the input coverage by how well the *translations* fit together to form a complete translation.

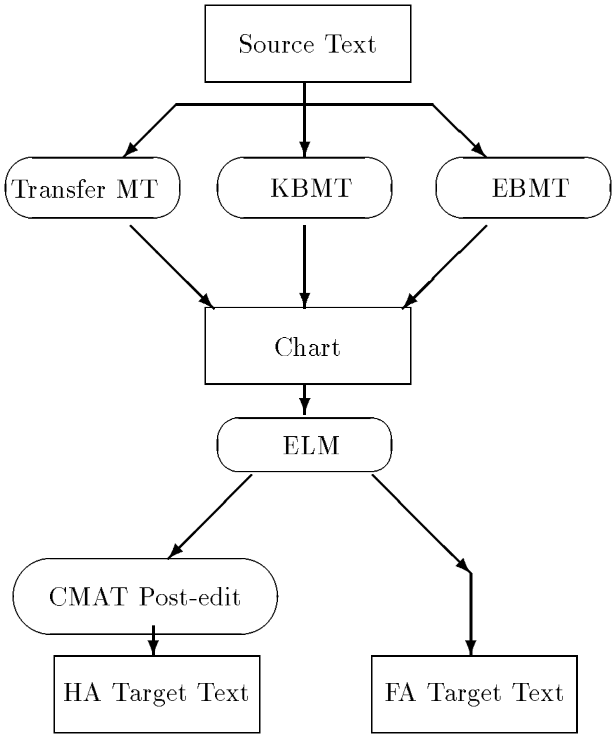

Figure 3: Pangloss Machine-Translation System Architecture

To find chunks, the engine sequentially looks up each word of the input in the index. The occurrence list for each word is compared against the occurrence list for the prior word and against the list of chunks extending to the prior word. For each occurrence which is adjacent to an occurrence of the prior word, a new chunk is created or an existing chunk is extended as appropriate. After processing all input words in this manner, the engine has determined all possible substrings of the input containing at least two words which are present in the corpus. Since the more frequent word sequences can occur hundreds of times in the corpus, the list of chunks is culled to eliminate all but the last five (by default) occurrences of any distinct word sequence. By selecting the *last* occurrences of each word sequence, one effectively gives the most recent additions to the corpus the highest weight, precisely what is needed for a translation memory.

Next, the sentence pairs containing the chunks found in the first phase are read from disk, and alignment is performed on each in order to determine the translation of the chunk–unless the match is against the entire corpus entry, in which case the entire target-language sentence is taken as the translation. Alignment currently uses a rather simplistic brute-force approach–very similar to that of (Nirenburg et al., 1994)–which identifies the minimum and maximum possible segments of the target-language sentence which could possibly correspond to the chunk, and then applies a scoring function to every possible substring of the maximum segment containing at least the minimum segment. The substring with the best score is then selected as the aligned match for the chunk.

The alignment scoring function is computed from the weighted sum of a number of extremely simple test functions. The weights can be changed for differing lengths of the source chunk in order to adapt to varying impacts of the tests with varying numbers of words in the chunk, as well as varying impacts as some or all of the raw test scores change. The test functions include (in approximate order of importance) such measures as **a)** the number of source words without correspondences in the target, **b)** the number of target words without correspondences in the source, **c)** matching words in source/target without correspondences, **d)** number of words with correspondence in the full target but not the candidate chunk, **e)** common sentence boundaries, **f)** elidable source words, **g)** insertable target words, and **h)** the difference in length between source and target chunks.

There is one exception to the above procedure for retrieving and aligning chunks. If any of the chunks covers the entire input string *and* the entire source-language half of a corpus sentence pair, then all other chunks are discarded and the target-language half of the pair is produced as the translation. This speeds up the system when operating in translation memory mode, as would be the case in a system used to translate revisions of previous texts. Unlike a pure translation memory, however, PanEBMT does not require an exact match with a memorized translation.

Figure 4 shows the set of translations generated from one sentence. The output is shown in the format used for standalone testing, which generates only the best translation for each distinct chunk; when integrated with the rest of Pangloss, PanEBMT also includes information indicating which portion of the input sentence and which pair from the corpus were used, and can produce multiple translations for each chunk. The number next to the source-language chunk in the output indicates the value of the scoring function, where higher values are worse. Very poor alignments (scores greater than five times the source chunk length) have already been omitted from the output.

6 Recent Enhancements

The EBMT engine described here is a completely new implementation in C++ replacing an earlier Lisp version. The previous version had performed very poorly (to the point where its results were

```
 El Banco de Santander habia sido
 elegido el lunes por las autoridades
 monetarias espanolas para comprar el
 Banco Espanol de Credito (Banesto),
 cuarto banco espanol.
"El Banco de" (0)
  ("the Bank of")
"El Banco de Santander" (1)
  ("the Bank of Santander")
"Banco de" (0)
  ("Bank of")
"Banco de Santander" (1)
  ("Bank of Santander")
"de Santander" (0)
  ("of Santander")
"habia sido" (0.5)
  ("been")
"elegido el" (0)
  ("chosen the")
"el lunes por" (0)
  ("Monday by the")
"por las" (0)
  ("by the")
"por las autoridades" (14.2)
  ("by the health authorities")
"por las autoridades monetarias" (0)
  ("by the monetary authorities")
"las autoridades monetarias" (0)
  ("the monetary authorities")
"comprar el" (0)
  ("buying the")
"Espanol de Credito" (13.2)
  ("Spanish Institute of Credit for")
"de Credito" (0)
  ("of credit")
"de Credito (" (1)
  ("of credit (")
"Credito (" (0)
  ("credit (")
", cuarto" (0)
  (", fourth")
"banco espanol" (0)
  ("Spanish bank")
"espanol ." (0)
  ("Spanish .")
```

Figure 4: Sample Translations

Input words		9169
Matched against corpus	90.4%	8294
Alignable	84.5%	7748
Good alignments	70.2%	6439

Table 1: Coverage and Sentence Alignability

Engine Name	Proposed Arcs	Proposed Words	Selected Arcs	Selected Words	Cover
DICT	27482	27482	3451	3451	9167
EBMT	11005	34992	1527	4768	6439
GLOSS	17663	19249	1567	1774	5780
Overall:	46580	71998	5415	9169	9169

Table 2: Contributions of Pangloss Engines

essentially ignored when combining the outputs of the various translation engines), for two main reasons: inadequate corpus size and incomplete indexing.

The earlier incarnation had used a corpus of considerably less than 40 megabytes of text, compared to the 270 megabytes used for the results described herein. The seven-fold increase in corpus size produces a proportional increase in matches.

Not only was the corpus fairly small, the text which was used was not fully indexed. To limit the size of the index file, a long list of the most frequent words were omitted from the index, as were punctuation marks. Although allowances were made for the words on the stop list, the missing punctuation marks always forced a break in chunks, frequently limiting the size of chunks which could be found. Further, allowance was made for the un-indexed frequent words by permitting *any* sequence of frequent words between two indexed words, producing many erroneous matches.

The newer implementation fully indexes the corpus, and thus examines only exact matches with the input, ensuring that only good matches are actually processed. Further, PanEBMT can index certain word pairs to, in effect, precompute some two-word chunks. When applied to the five to ten most frequent words, this pairing can reduce processing time during translation by dramatically reducing the amount of data which must be read from the index file (for example, there might be 10,000 occurrences of a word pair instead of 1,000,000 occurrences of one of the words and 100,000 of the other word), and thus the number of adjacency comparisons which must be made.

7 Performance

7.1 Accuracy

PanEBMT was first put to the test during an internal evaluation in August 1995, which was similar in design to the ARPA MT evaluations (White & O'Connell, 1994). During this evaluation, twenty newswire articles (selected from the 100 articles used in the prior ARPA evaluation) averaging about 450 words each were processed and subsequently examined. For this paper, another evaluation was performed using a subset of the Pangloss system on the 253 sentences in the twenty articles. Table 2 shows the total number of arcs proposed by each translation engine used, the number selected for output by the statistical language model, and the number of source words represented by those arcs. The final column shows the total number of source words covered by at least one proposed arc. The values for individual engines do not sum to the *Overall* value because multiple engines can produce equivalent arcs, which are combined in the chart, with both engines credited for the arc. The engines listed in the tables are

- **DICT**ionary: PanEBMT's association dictionary, used here primarily to provide coverage for words not otherwise covered
- **EBMT**: PanEBMT
- **GLOSS**aries: hand-crafted word/phrase bilingual glossaries

7.2 Speed

Indexing a 270 megabyte corpus requires approximately 45 minutes on a Sun Sparcstation LX when all files are located on local disks, and another 80 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.

A sample text of 15 sentences totalling 414 words and punctuation marks can be processed in just under three minutes. The 20 texts used in the evaluation can be completely processed in two hours, including separate passes for dictionary lookups and statistical modeling by a separate program (described in (Brown and Frederking, 1995)); PanEBMT accounts for about 80 minutes of those two hours.

The above timings represent a variety of speed optimizations which have been applied since the August 1995 evaluation, resulting in a doubling of the indexing speed and tripling of translation speed.

8 Strengths and Weaknesses

As currently implemented, PanEBMT has both strengths and weaknesses. Its strengths are that the minimal knowledge required allows quick retargeting and that its design provides for graceful degradation. Its weaknesses are that it is unable to completely cover inputs, that it does not perform well when the correspondences between source-language and target-language words are not one-to-one, and that (like statistically-based translation systems) it is sensitive to differences between the example corpus and the sentences to be translated.

The astute reader will have noticed that there have been virtually no mentions of the source or target languages in this paper–they are not relevant to discussions of the design and operation of the engine, since the only language-dependent knowledge consists of the equivalence classes and the lists of insertable and elidable words, which are provided via the configuration file. This language-independent aspect of EBMT makes PanEBMT rapidly retargetable to other language pairs, and in fact there are already versions of PanEBMT providing Serbocroatian-to-English and English-to-Serbocroatian translations (no experimental data is as yet available for Serbocroatian because the complete dictionary and corpus are still being acquired). Given the three required knowledge sources of corpus, dictionary, and word-root list, PanEBMT can begin producing translations for a new language pair in only a few hours. Fine tuning will require one to two weeks to determine reasonable word classes for tokenization (along with the required re-indexing of the corpus) and to adjust the scoring function weights.

Number and quality of translations degrades gradually as the size and quality of the bilingual dictionary and synonym list decrease. An incomplete dictionary or root/synonym list merely causes PanEBMT to miss some potential translations. Similarly, a smaller corpus produces fewer potential matches, but there is no point for any of the three knowledge sources at which the engine suddenly ceases to function. One can take advantage of this gradual behavior by building the knowledge sources incrementally and using EBMT for translations even before the knowledge sources have been completed. In particular, by adding post-edited output of the MT system back into the corpus, the system can be bootstrapped from a relatively modest initial corpus (precisely the idea behind a translation memory).

During preparation of this paper, several extraneous lines were discovered in the corpus files, which caused more than 29,400 sentence pairs (over 4% of the corpus) to be corrupted. Due to the extra lines, the corrupted pairs consisted of the English target sentence from one pair and the Spanish source sentence from the following pair. This error had not been discovered earlier because it had no obvious effect on PanEBMT's performance – a clear example of the system's graceful degradation property.

Lack of complete input coverage is a severe obstacle to using PanEBMT as a stand-alone translation system. The engine can not generate a chunk for a word unless it both co-occurs with either the preceding or following word somewhere in the corpus, and at least one occurrence can be successfully aligned. Additionally, candidate chunks are omitted if the alignment was successful but the scoring function indicates a poor match. Unless all of these conditions are met, a gap in output occurs for the particular input word. In the context of the Pangloss system, such gaps are not a problem, since one of the other engines can usually supply a translation covering each gap.

As currently implemented, the EBMT engine is unable to properly deal with translations that do not involve one-for-one correspondences between source and target words (e.g. Spanish "mil milliones" corresponding to English "billions"). Lack of a one-to-one correspondence between source-language and target-language expressions can often cause the alignment to be incorrect or fail altogether under the current alignment algorithm.

Since the corpus used in the experiments described here was based almost entirely on the UN proceedings rather than newswire text, PanEBMT did not find many long chunks during the evaluation. In fact, the average chunk was just over three words in length, and less than three percent of the chunks were more than six words long. This quite naturally affects the quality of the final translation, since many short pieces must be assembled into a translation rather than one or two long segments.

Despite all these difficulties, PanEBMT was able to cover 70.2% of the input it was presented with good chunks, and generate some translation for more than 84ordinarily not output at all). Integrating the hand-crafted glossaries from Pangloss into the corpus, thus adding 148,600 effectively pre-aligned phrases to the corpus, improved the matches against the corpus from 90.4% to 90.9% of the input, and the coverage with good chunks to 73.3%.

9 Future Enhancements

Since PanEBMT is a fairly new implementation, there is still much that could be done to enhance it. Among the improvements being considered are: improving the quality of the dictionary (in progress); supporting one-to-many or many-to-one associations for alignment; optimizing the test-function weights; other alignment algorithms; using linguistic information such as morphological variants and source-language synonymy to increase the number of matches against the corpus; using approximate matchings when no exact matches exist in the corpus; and using of a classifier algorithm to remove redundancy from the corpus (suggested by C. Domashnev).

References

Ralf Brown (in preparation). FramepaC User's Manual Carnegie Mellon University Center for Machine Translation technical memorandum *http://www.cs.cmu.edu/afs/cs.cmu.edu/-user/ralf/pub/WWW/papers.html*

Ralf Brown and Robert Frederking 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), pages 221–239. Leuven, Belgium.

David Graff and Rebecca Finch 1994. Multilingual Text Resources at the Linguistic Data Consortium In *Proceedings of the 1994 ARPA Human Language Technology Workshop* Morgan Kaufmann.

H. Maruyama and H. Watanabe 1992. Tree Cover Search Algorithm for Example-Based Translation. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), pages 173–184. Montreal.

M. Nagao 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In *Artificial and Human Intelligence*, A. Elithorn and R. Banerji (eds). NATO Publications

Sergei Nirenburg, (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.

Sergei Nirenburg, Stephen Beale, and Constantine Domashnev 1994. A Full-Text Experiment in Example-Based Machine Translation. In *New Methods in Language Processing* Manchester, England.

Sergei Nirenburg, Constantine Domashnev, and Dean J. Grannes 1993. Two Approaches to Matching in EBMT. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-93).

White, J.S. and T. O'Connell. 1994. "Evaluation in the ARPA Machine Translation Program: 1993 Methodology." In *Proceedings of the ARPA HLT Workshop*. Plainsboro, NJ.