Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation

Ralf D. Brown

Language Technologies Institute
(Center for Machine Translation)
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA
ralf+@cs.cmu.edu

Abstract. An Example-Based Machine Translation system is supplied with a sentence-aligned bilingual corpus, but no other knowledge sources. Using the knowledge implicit in the corpus, it generates a bilingual word-for-word dictionary for alignment during translation. With such an automatically-generated dictionary, the system covers (with equivalent quality) *more* of its input on unseen texts than the same system does when provided with a manually-created general-purpose dictionary *and* other knowledge sources.

1 Introduction

Previous work ([Brown, 1996, Frederking and Brown, 1996]) on the Pangloss Example-Based Machine Translation engine (PanEBMT) has always assumed the availability of knowledge sources in addition to the sentence-aligned bilingual corpus, particularly a *large* bilingual dictionary. Although more readily available and/or acquired than, for example, the ontologies and other knowledge sources for a knowledge-based translation system, generating these additional EBMT knowledge sources manually still represents a considerable investment of effort. Acquiring and correcting a 62,000-entry Croatian dictionary for the DIPLOMAT project ([Frederking *et al*, 1997]) required about eight person-months of effort.

Given that the aligned bilingual corpus represents a considerable amount of implicit knowledge, a means of extracting that implicit knowledge into the knowledge sources required by PanEBMT would greatly speed development of new language pairs. Fortunately, only one knowledge source other than the corpus itself is absolutely required for PanEBMT to operate: a large bilingual dictionary (see also Table 1). Statistical MT work such as that of [P. Brown *et al*, 1988] demonstrated many years ago that it is possible to generate a corpus-based dictionary with a considerable degree of accuracy, so an obvious improvement to PanEBMT was to have it create its own dictionary from its corpus.

This paper describes the results of experiments using various dictionaries created from a Spanish-English corpus consisting of some 685,000 sentence pairs derived primarily from the Spanish and English portions of the UN Multilingual Corpus ([Graff and Finch, 1994]). The dictionaries were used to replace a 51,500-entry dictionary derived from the Collins Spanish-English dictionary and a 26,500-entry English root/synonym list derived from WordNet (the latter was added to compensate for the fact that most translations in the Collins-based dictionary are uninflected and thus often fail to match the surface forms found in the corpus).

A major advantage of a corpus-derived dictionary over a general-purpose dictionary is that it is tuned to the way the corpus translates its sentences. For example, the UN corpus translates "alertas" as "warnings" much more frequently than as "alerts", yet neither the Collins dictionary

Table 1. Language-Specific Knowledge Sources

Knowledge Source	Standard Operation	"Knowledge-Free"
parallel corpus	UN corpus	UN corpus
dictionary	Collins	extracted from corpus
root/synonym list	WordNet-derived	none
tokenizations	as indicated (see text)	as indicated (see text)
elidable-words	"se", "su", "el",	"se", "su", "el",
	"una", "de", "del",	"una", "de", "del",
	"la", "los", "las",	"la", "los", "las",
insertable-words	"a", "an", "the", "of"	"a", "an", "the", "of"

nor the WordNet-derived English synonym list allow PanEBMT to determine that "warnings" is a possible translation for "alertas" – it is neither one of the translations given by Collins nor in the synonym lists of any of the given translations. In contrast, the corpus-derived dictionary not only lists both "warnings" and "alerts" as translations, it can also record that the former is some four times more likely (though that fact is not yet used).

2 Extracting a Bilingual Dictionary

The dictionary needed for PanEBMT to perform subsentential alignment is extracted from the corpus using a correspondence table which is filtered using a thresholding scheme (rather than a measure such as mutual information or Dice coefficients). Any word pairs which pass the threshold filter are considered to be translations for the purposes of EBMT alignment.

The correspondence table is a two-dimensional array indexed by source-language words in one dimension and target-language words in the other. For each sentence pair in the corpus, all entries corresponding to the cross-product of the source-language sentence and target-language sentence (after removing duplicate words in each sentence) are incremented. In addition, the monolingual occurrence counts for each of the unique source and target words are incremented for use in the filtering phase.

For language pairs which have generally similar word orders, word pairs in roughly equivalent positions within each sentence can have their correspondence-table entries incremented twice, in order to bias the correspondence table toward portions of the target sentence which are most likely to be the translation of that portion of the source sentence.[1]

Once all of the sentence pairs in the corpus have been processed, the correspondence table is filtered using a symmetric co-occurrence ratio and an asymmetric co-occurrence ratio, both of which may vary as a function of the total co-occurrence count. Two distinct variations in threshold-setting were investigated: a step function which sets the thresholds to an unreachably high value for co-occurrence counts less than some minimum (2 in the described experiments) and to a constant value in all other cases; and a sliding scale starting at 1.0 for a co-occurrence count of 1, decreasing smoothly to some minimum threshold value, in order to reduce the number of coincidental co-occurrences which pass the filtering. Any elements of the table which fail both ratio tests are set to zero. All remaining non-zero elements are then added to the dictionary, along with their co-occurrence counts.

[1] A variant of this refinement was suggested by Christopher Hogan.

The two co-occurrence ratio tests are used to determine whether a target-language word is found in the translation of a sentence containing the source-language word sufficiently frequently to be a probable translation of the source language word. The symmetric threshold is passed whenever

$$C[S,T] \geq threshold[C] * count[S] \text{ and } C[S,T] \geq threshold[C] * count[T],$$

where $C[S,T]$ is the number of times source-language word S co-occurs with target-language word T and $threshold[C]$ is the threshold value selected by that co-occurrence count[2]. The asymmetric threshold is passed whenever

$$C[S,T] \geq thresh1[C] * count[S] \text{ and } C[S,T] \geq thresh2[C] * count[T], \text{ or}$$
$$C[S,T] \geq thresh1[C] * count[T] \text{ and } C[S,T] \geq thresh2[C] * count[S],$$

where $thresh1[C]$ and $thresh2[C]$ are the two separate limits of the asymmetric threshold. This second test is used to account for words which are polysemic in one language but not the other.

Despite the simplicity of the above algorithm, it performs quite well. By setting the thresholds used in filtering to different values, a tradeoff between yield and accuracy may be tuned (see Table 2); note that the error rate is based on total definitions, and that a far lower percentage of words have *only* incorrect definitions. Raising the thresholds reduces the number of incorrect/spurious translations generated, but reduces the size of the dictionary (e.g. 72% precision at 29% recall). Lowering the thresholds yields more definitions, but also increases the error rate (e.g. 46% precision at 45% recall). The yields and accuracies shown here are based on merging two partial dictionaries generated from a 60/40 corpus split (that being a convenient place); the dictionaries generated by treating the corpus as a monolithic whole not only contained more spurious translations, but also had a slightly *smaller* vocabulary. As will be shown, the alignment algorithm used by PanEBMT is robust enough to tolerate a significant number of erroneous translations, particularly when most of the errors are in the lower-frequency terms, as is the case with the automatically-extracted dictionaries.

The final dictionary listed in the left-hand column of Table 2 was created by first generating another dictionary consisting entirely of words occurring only a single time which correspond to target-language words which also have a single occurrence. The resultant dictionary of 15,979 singleton terms with an error rate of 25% was then merged with the dictionary previously created using a fixed filtering threshold of 0.10.

Once the biasing option was implemented, an additional set of dictionaries was created. In generating the co-occurence table for these dictionaries, double weight was given to words within the "expected" range in the target-language sentence, computed as follows: Treat the source sentence as the interval [0.0,1.0] and determine the source word's location, i.e. 0.63. Find the word in the target-language sentence at the equivalent location, and then expand the range to include +/- 0.15 of the sentence, but no less than two words in each direction, from that word (e.g. the interval [0.48,0.78]). Multiple dictionaries with the same bias and minimum threshold were generated, differing in how quickly the minimum threshold was reached with increasing co-occurrence counts; these are identified by letter, where higher letters indicate that the minimum is reached more quickly.

In addition to a high error rate on word pairs with low co-occurrence counts (as one would expect), there are two major cases in which the co-occurrence dictionary consistently generates erroneous translations. These involve the very highest-frequency words, and words which typically co-occur monolingually. The highest-frequency words occur in so many sentences that they

[2] When the source and target words occur with equal frequency, this threshold test becomes equivalent to the Dice coefficient used by [Kitamura and Matsumoto, 1996].

Table 2. Accuracy vs. Coverage

Fixed Thresholds	Vocabulary Size	Estimated Error Rate	Variable Thresholds	Vocabulary Size	Estimated Error Rate
entire corpus	96,793	—			
1.00	14,446	29%	0.10 a	28,193	28%
0.40, 0.33/0.60	24,034	31%	0.10 b	36,447	38%
0.33, 0.25/0.60	26,060	34%	0.05 a	33,446	37%
0.25, 0.20/0.50	28,543	38%	0.05 b	40,632	49%
0.20, 0.15/0.50	30,871	41%	0.05 c	42,006	52%
0.15, 0.12/0.50	33,482	43%	0.05 d	42,868	53%
0.12, 0.10/0.50	35,444	45%	0.05 e	43,409	54%
0.10	36,854	46%			
0.10+singletons	52,833	40%			

perforce co-occur with many unrelated high-frequency words in the other language sufficiently often to pass the threshold tests. Similarly, monolingual co-occurrences such as in the country name "Burkina Faso" – which appears as such in both sides of the corpus – will generate individual translations for each of the words of its translation (e.g. both "Burkina" and "Faso" will list translations of both "Burkina" and "Faso").

Initial experiments indicated that the error-ridden high-frequency terms added an unacceptable amount of noise to the correspondence table used for alignment, so a second pass of dictionary extraction is performed with a slightly modified algorithm. Given a list of the highest-frequency terms in the corpus (in this case, all words which appear in at least 20% of the source sentences), all sentence pairs are skipped except those containing exactly one or two of the high-frequency words. This permits better discrimination between the translations of the individual frequent words, resulting in a secondary dictionary containing 7 of the 16 high-frequency words, with a zero error rate. This secondary dictionary is then merged with the result of the first pass, such that entries in the secondary dictionary override those in the main dictionary.

3 Matching Inputs with the EBMT Corpus

Unlike most other EBMT systems, including some early experiments in the Pangloss project ([Nirenburg *et al*, 1993]), PanEBMT does not find the corpus sentence which most nearly matches the input and then modify the translation, but rather finds all matching substrings of the input in the corpus, and then attempts to identify the translation of each match within the full sentence pair. Each partial translation is output, to be combined by the translations system in a chart with results from other engines and eventually given to a statistical language modeler ([Brown and Frederking, 1995]) for selection of the final translation.

When performing translations, PanEBMT uses an inverted index built from the source-language half of the bilingual corpus. Matches are found by consulting the index to identify adjacent occurrences in the corpus of words which are also adjacent in the input; each match is extended to cover as much of the input as possible. The last N (usually 8) occurrences of any particular substring of the input are used to find a translation of that substring; the remaining occurrences are discarded to avoid excessive processing for high-frequency phrases (several two-word phrases occur more than 100,000 times each in the UN corpus).

4 Subsentential Alignment

Once matching phrases in the corpus have been found, the individual sentence pairs containing the matches are retrieved from the corpus, and subsentential alignment is performed to determine the translation of the matched portion. Alignment consists of two main phases: generating a possible-translation table[3] and applying a set of heuristic scoring functions to substrings of the target-language sentence using that table.

A correspondence table (which is simply an incomplete, ambiguous bitext map) for a sentence pair is built by looking up the translation of each word in the source half as well as the synonym list for each word in the target half. A word pair (S,T) is marked as corresponding (and thus a possible translation) if:

1. T is identical to S or appears in the list of translations for S
2. the list of translations for S and the list of synonyms for T have any members in common.

For language pairs with similar word orders (such as Spanish and English), the initial correspondence table is further pruned by removing outliers. For each word triple, the earliest and latest possible positions of the first and third words in the other language are determined, and this range is expanded by N (in this case, 2) words on both the left and right to allow for word-order variations. If the second word of the triple has correspondences both within and outside these limits, the possible correspondences which lie outside the limits are erased.

After the correspondence table has been built, it is searched for one or more "anchors" within the matched input segment. An anchor is a word which uniquely corresponds between source and target languages – it has only one possible translation listed, and is the only known translation for its translation. If no anchors are found within the matched segment, or multiple anchors whose target-language translations are deemed to be too far apart, the sentence is considered unalignable and processing skips to the next corpus match. Otherwise, the target-language substring containing all anchors is marked as the minimum possible translation. The substring containing the minimum translation plus all left- and right-adjacent words not known to translate only words outside the matched input is marked as the maximum possible translation. The heuristic scoring functions (which include lists of words which may be elided or inserted without the full penalty for an unmatched word; see Table 1) are then applied to all substrings of the maximal translation which include at least the minimum translation, and the substring with the best score is output as the translation of the matched phrase. Removing the outliers from the correspondence table improves the alignment process both by creating more anchors and by removing some spurious correspondences which produce a larger maximal translation.

5 Experimental Setup

The performance of PanEBMT on two different test texts was measured with each of several different automatically-generated dictionaries, as well as the previously-used configuration of the Collins dictionary in conjunction with other knowledge sources. Additionally, each text was tested both with and without the manually-created tokenization file, which is used to increase the number of matches against the corpus. This is a small file which lists 47 equivalence classes (e.g. conjunctions "and" and "or", month names, country names, and days of the week)

[3] This table may be precomputed when the corpus is indexed for faster translations.

containing a total of 880 words and the translation of each word; it represents about one day's effort for a person fluent in both languages. The tokenization file is used during corpus matching (and for subsequent back-substitution of the appropriate term), but does not augment the dictionary during subsentential alignment.

The two test texts contain 275 sentences of the UN corpus which were omitted from the EBMT corpus and 253 sentences of newswire text from a 1994 ARPA MT evaluation. These texts may be considered in-domain and out-of-domain; the UN sentences are very similar to the remainder of the corpus, and thus produce more and larger matches against the corpus than the out-of-domain newswire sentences.

Each of the four configurations (UN text with and without tokenization file, and newswire text with and without tokenization file) was tested using each corpus-derived dictionary, the Collins-derived dictionary, and the Collins-derived dictionary in conjunction with the WordNet-derived root/synonym list.

6 Results

The results of the evaluation are shown in Table 3. The first line indicates the proportion of the input texts for which multi-word matches could be found in the parallel corpus. This value is the maximum coverage that PanEBMT could achieve, given perfect alignments for all matches. In practice, the actual coverage is lower both because alignment sometimes fails, and because the best alignments of some sentences have such poor scores that they are discarded. The remainder of Table 3 indicates how PanEBMT peformed on each text with each of the different dictionaries; the dictionaries are identified by the symmetric threshold used in their generation (step function), the minimum threshold (smooth threshold function) together with an identifiation letter, or as "Coll+WN" and "Collins" for the Collins-based dictionary with/without the WordNet-derived synonym list.

As can be seen from Table 3, the best of the corpus-derived dictionaries (without using the tokenization file) had a coverage which was 106% of the manual-dictionary performance (even *using* the tokenization file) on the UN text, and 99% of the best manual-dictionary performance on the Spanish newswire text, while maintaining the same level of quality. Performance is even better if the corpus-derived dictionaries are allowed to make use of tokenizations.

Table 3 clearly shows that the PanEBMT alignment process needs a *large* bilingual dictionary – the increased error rate from using a lower threshold during dictionary extraction is more than outweighed by the increased vocabulary. The alignment process can tolerate a high dictionary error rate in part because most segments of the input have multiple matches in the corpus, only one of which must have a good alignment for translation to succeed. Further, erroneous definitions for low-frequency words can be tolerated because it is unlikely that both a correct and incorrect definition will be matched in the same sentence.

Not surprisingly, the accuracy of the dictionary on the highest-frequency words is quite important. The initial runs with the "0.10" dictionary accidentally omitted the corrections generated by the high-frequency extraction algorithm, and resulted in a higher coverage than reported here, but a significantly worse quality. In practice, one would want to manually correct the top 50 to 100 words of the dictionary to ensure maximum performance with minimum effort.

In addition to their use in subsentential alignment for PanEBMT, the corpus-derived dictionaries have also been used in a series of experiments in Translingual Information Retrieval [Carbonell *et al*, 1997]. The best English-to-Spanish dictionary derived from the corpus (using a sliding threshold with a minimum value of 0.27) outperformed in absolute precision the

Table 3. EBMT Coverage

Dictionary	UN text no tokens	UN text w/ tokens	newswire no tokens	newswire w/ tokens
Corpus	98.0%	98.1%	90.1%	90.5%
Collins	48.0%	49.0%	25.1%	27.0%
Coll+WN	82.7%	83.3%	73.8%	75.3%
(Fixed Thresholds)				
0.40	65.9%	66.8%	55.0%	56.6%
0.33	69.5%	70.1%	58.2%	59.6%
0.25	75.7%	76.2%	63.1%	64.5%
0.20	80.0%	80.4%	65.7%	67.2%
0.15	82.1%	82.5%	68.0%	69.4%
0.12	84.1%	84.6%	69.5%	70.7%
0.10+sing	84.9%	not run	70.3%	not run
(Variable Thresholds)				
0.10 a	62.5%	63.2%	52.5%	53.7%
0.10 b	73.6%	74.2%	60.7%	61.9%
0.05 b	77.1%	77.7%	64.1%	65.2%
0.05 c	82.0%	82.5%	69.3%	70.4%
0.05 d	83.7%	84.3%	70.8%	71.9%
0.05 e	88.3%	88.6%	74.7%	75.6%

other translingual methods which were investigated, and was only slightly worse in relative Translingual/Monolingual performance.

7 Future Enhancements

Use of corpus-derived dictionaries for PanEBMT is a recent development, and a number of obvious enhancements are still to be made:

- identification of phrases in the dictionary, e.g. using the methods described in [Wu, 1995] or [Kaji and Aizono, 1996]. PanEBMT's alignment code can take advantage of dictionary entries which have multi-word translations, such as are already present in the Collins-derived dictionary.
- iterative refinement of the dictionary – the alignment process may identify additional translations, and comparing the actual translations used in alignment may help remove spurious entries from the dictionary.
- improved subsentential alignment, culminating in complete word-level alignment of the entire bilingual corpus. This improvement will show synergy in conjunction with the iterative dictionary refinement.

Further opportunities for improvement exist in tuning the initial dictionary extraction itself. Additional experiments are required to determine the optimum number of parts into which to split the corpus to generate partial dictionaries. Increasing the number of parts will reduce the chance of a coincidental co-occurrence being accepted as a translation, but will also reduce the yield as lower-frequency terms become too infrequent.

8 Acknowledgements

This work was supported as part of the DIPLOMAT project, funded under DARPA grant number N000 149312005 N203B.

References

[P. Brown *et al*, 1988] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A Statistical Approach to Language Translation. In *COLING-88: The 13th International Conference on Computational Linguistics*, Budapest, pp. 71–76.

[Brown, 1996] Ralf D. Brown. Example-Based Machine Translation in the Pangloss System. In *COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, pp. 169–174.

[Brown and Frederking, 1995] Ralf D. Brown and Robert Frederking. 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), Leuven, Belgium, pp. 221–239.

[Carbonell *et al*, 1997] Jaime Carbonell, Yiming Yang, Robert Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual Information Retrieval: A Comparative Evaluation. To appear in *Proceedings of Fifteenth International Joint Conference on Artificial Intelligence* (IJCAI-97).

[Frederking and Brown, 1996] Robert Frederking and Ralf D. Brown. The Pangloss-Lite Machine Translation System. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, pp. 268–272.

[Frederking *et al*, 1997] Robert Frederking, Alexander Rudnicky, and Christopher Hogan. Interactive Speech Translation in the DIPLOMAT Project. *Workshop on Spoken Language Translation at ACL-97*, Madrid.

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

[Kaji and Aizono, 1996] H. Kaji and T. Aizono. Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information. In *COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, pp. 23–28.

[Kitamura and Matsumoto, 1996] M. Kitamura and Y. Matsumoto. Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pp. 79–87. Copenhagen, Denmark.

[Maruyama and Watanabe, 1992] H. Maruyama and H. Watanabe. Tree Cover Search Algorithm for Example-Based Translation. In *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT*, Montreal, pp. 173–184.

[Nagao, 1984] M. Nagao. 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

[Nirenburg *et al*, 1994] Sergei Nirenburg, Stephen Beale, and Constantine Domashnev. A Full-Text Experiment in Example-Based Machine Translation. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, pp. 78–87.

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

[Wu, 1995] D. Wu. Grammarless Extraction of Phrasal Translation Examples from Parallel Texts. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, pp. 354–372.