A Compact Data Structure for Searchable Translation Memories

Chris Callison-Burch **Colin Bannard**
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
{chris,colin}@linearb.co.uk

Josh Schroeder
Linear B Ltd.
39 B Cumberland Street
Edinburgh EH3 6RA
josh@linearb.co.uk

Abstract

In this paper we describe searchable translation memories, which allow translators to search their archives for possible translations of phrases. We describe how statistical machine translation can be used to align subsentential units in a translation memory, and rank them by their probability. We detail a data structure that allows for memory-efficient storage of the index. We evaluate the accuracy of translations retrieved from a searchable translation memory built from 50,000 sentence pairs, and find a precision of 86.6% for the top ranked translations.

1 Introduction

The work of any translator or translation agency contains significant amounts of repetition, and translation archives are consequently a vital asset. Current translation memory systems provide a valuable means for translators to exploit this resource in order to increase productivity and to ensure consistency. Existing translation memory systems work by retrieving the translation of full sentences that are exactly or approximately matched in a database of a translator's past work (Trujillo, 1999). Translation memories provide facilities for the automatic alignment of sentences and paragraph units (Gale and Church, 1993; Kay and Röscheisen, 1993), but aligning subsentential units is usually an involved, manual process.

Matching on the sentence-level is a rather severe restriction which means that only very limited reuse is made of the information contained within a translation archive. A translator will frequently use phrases, words and other subsentential strings that s/he has translated before. However, unless these are contained as a whole unit within the database, conventional translation memory systems are unable to retrieve translations for them.

This paper describes a search tool which allows more flexible information retrieval than sentence-level matching. The usefulness of a translation database might be greatly increased if it could be easily searched, for example by returning focused translations when a user queries it with a single phrase. This paper describes tools which offer precisely that facility. We present *searchable translation memories* which allow Google-style searching of translation archives.

Figure 1 illustrates the use of the technology. The figure shows example results of querying a searchable translation memory built from French and English portions of the proceedings of the European Parliament. The user has typed the search phrase *west bank*, and similar to a parallel concordancer (Barlow, 2004), the system has returned a list of sentences that the phrase occurs in. However, unlike a concordancer, the searchable translation memory picks out those phrases which constitute the likely translations of the phrase (*cisjordanie, territoires de cisjordanie, rive ouest,* and *rive gauche du jourdain*), groups retrieved sentences by these translations, and ranks the groups according to their probability.

There are two primary technical challenges for searchable translation memories. The first is the ability to index a translation memory so that it contains the correspondences between translated words and phrases across the two languages. For this we

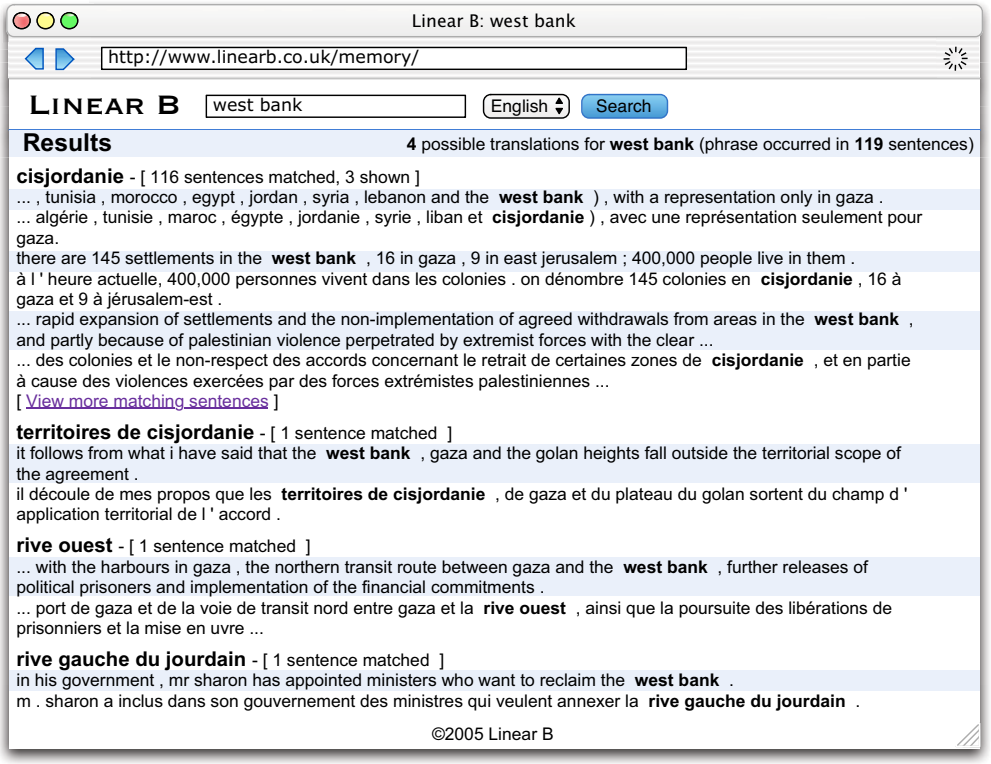

Figure 1: Search results for the English phrase “west bank”

rely on phrase-based statistical machine translation (Koehn et al., 2003). The second is the efficient storage of the index, for which we use a variant of the suffix array data structure (Manber and Myers, 1990).

In this paper we:

- Review the concepts in statistical machine translation which are relevant to the alignment of phrases
- Demonstrate how a word- and phrase-aligned parallel corpus can be efficiently indexed for searching with suffix arrays
- Evaluate the accuracy of the phrases retrieved by our system using the precision and recall metrics that are standard in information retrieval
- Discuss why tools such as these may ultimately be more useful than fully-automatic machine translation

2 Statistical Machine Translation

Usually the goal of statistical machine translation (Brown et al., 1988) is to be able to choose that target language (English) sentence, $\mathbf{e}$, that is the most probable translation of a given sentence, $\mathbf{f}$, in a foreign language. Rather than choosing $\mathbf{e}^*$ that directly maximizes the conditional probability $p(\mathbf{e}|\mathbf{f})$, Bayes' rule is generally applied:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} p(\mathbf{e})p(\mathbf{f}|\mathbf{e}) \quad (1)$$

In this equation $p(\mathbf{e})$ is a language model probability of the translation and $p(\mathbf{f}|\mathbf{e})$ is a translation model describing the stochastic mapping of a source sentence onto a target sentence. The effect of applying Bayes' rule is to divide the task into estimating two probabilities: a language model probability $p(\mathbf{e})$ which can be estimated using a monolingual corpus, and a translation model probability $p(\mathbf{f}|\mathbf{e})$ which is estimated using a bilingual sentence-aligned corpus, such as a translation memory.

The next two subsections examine how the trans-

lation model probability is calculated using subsentential alignments. We use these alignments to construct a searchable translation memory, rather than for the task of fully automatic machine translation.

2.1 Word Alignments

Brown et al. (1993) formulate translation essentially as a word-level operation. The probability that a foreign sentence is the translation of an English sentence is calculated by summing over the probabilities of all permissible word-level alignments, $\mathbf{a}$, between the sentences:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad (2)$$

where an alignment $\mathbf{a}$ is defined as a subset of the Cartesian product of the word positions in $\mathbf{e}$ of length I and $\mathbf{f}$ of length J:

$$\mathbf{a} \subseteq \{(i, j) : i = 1...I; j = 1...J\} \quad (3)$$

Thus Brown et al. decompose the problem of determining whether a sentence is a good translation of another into the problem of determining whether there is a sensible mapping between the words in the sentences.[1] Figure 2 illustrates a probable word-level alignment between a sentence pair in the Canadian Hansard bilingual corpus.

2.2 Phrase Alignments

Phrase-based translation uses larger segments of human translated text. Phrase-based translation models provide an estimate of phrasal translation probabilities. The probability of an English phrase $\bar{e}$ translating as a foreign phrase $\bar{f}$ can be calculated using maximum likelihood estimation

$$p(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e})} \quad (4)$$

where counts are collected over each instance where $\bar{e}$ is aligned with $\bar{f}$ in any sentence pair in the training corpus. Prior to applying this probability assignment one must define a method for determining which phrases are aligned.

[1]For brevity, we have omitted the details about how the parameters of $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$ are estimated using expectation maximization, and instead refer the reader to Brown et al. (1993), Knight (1999) and Och and Ney (2003).

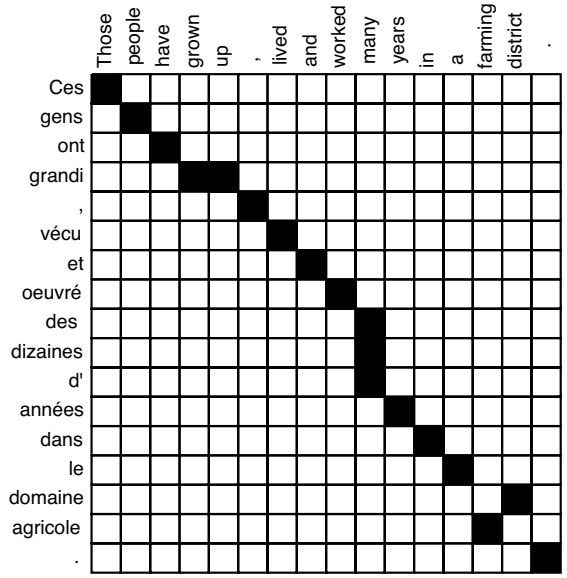

Figure 2: A word-level alignment for a sentence pair that occurs in our training data

There are various heuristics for extracting phrase alignments from word alignments,[2] some are described in Koehn (2004), Tillmann (2003), and Vogel et al. (2003). We define phrase alignments as follows. A substring $\bar{e}$ consisting of the words at positions $l...m$ is aligned with the phrase $\bar{f}$ by way of the subalignment $\mathbf{s}$

$$\mathbf{s} = \mathbf{a} \cap \{(i, j) : i = l...m, j = 1...J\} \quad (5)$$

$\bar{f}$ is the phrase corresponding to the words formed by ordering the set of indices j in (i, j) in $\mathbf{s}$. Note that the 'phrases' in phrase-based translation do not correspond to the traditional notion of syntactic constituents; they might be more aptly described as 'substrings' or 'blocks'.

Some examples of phrase alignments that can be extracted from Figure 2 include: *lived and worked → vécu et oeuvré, many years → des dizaines d'années, a farming district → le domaine agricole*. Strictly speaking our method for extracting phrase alignments does not require that $\bar{f}$ be a contiguous phrase. We insert an placeholder element to indicate

[2]There are other ways of calculating phrasal translation probabilities. For instance, Marcu and Wong (2002) estimate them directly rather than starting from word-level alignments.

any discontinuous span in $\bar{f}$. For example we retrieve the alignment *a farming* → *le ... agricole*. Our definition for phrase alignments is useful because it ensures that we are able to retrieve a possible translation for any phrase that occurred in the source corpus.

3 Constructing an Index for Searchable Translation Memories

Once we have defined a method for extracting phrase alignments, then we can construct an index for our searchable translation memories. This index will allow us to retrieve the translations of a search query, along with a set of sentences which illustrate the context in which the translations appear.

One way of indexing a translation memory would be to keep a table of all source phrases with their corresponding target phrase alignments, and a list of the indices of all the sentence pairs that a phrase alignment occurred in. If we created an index containing each phrase, its possible translations, and a link back to the original sentences then a record in our index would look like the following:

source	→	possible	sentence pairs it
phrase		translation 1	occurred in
	→	possible	sentence pairs it
		translation 2	occurred in

With such an index, it is a simple matter to query it with a certain source phrase, retrieve all possible translations and their contexts, and rank the translations using the phrase translation probability calculation in Equation 2.2.

However from an engineering perspective such an index would be unwieldy in terms of the amount of space needed to save an enumeration of all possible subphrases in the corpus. The problem of memory usage when enumerating phrases and their translations is described in detail in Callison-Burch et al. (2005). The next section describes our use of suffix arrays as an alternative to an enumerated index.

4 Suffix Arrays

The suffix array data structure (Manber and Myers, 1990) was introduced as a space-economical way of creating an index for string searches. The suffix array data structure makes it convenient to compute the frequency and location of any substring or n-gram in a large corpus. Abstractly, a suffix array is an alphabetically-sorted list of all suffixes in a corpus, where a suffix is a substring running from each position in the text to the end. However, rather than actually storing all suffixes, a suffix array can be constructed by creating a list of references to each of the suffixes in a corpus. Figure 3 shows how a suffix array is initialized for a corpus with one sentence. Each index of a word in the corpus has a corresponding place in the suffix array, which is identical in length to the corpus. Figure 4 shows the final state of the suffix array, which is as a list of the indices of words in the corpus that corresponds to an alphabetically sorted list of the suffixes. The advantages of this representation are that it is compact and easily searchable. The total size of the suffix array is a constant amount of memory.

Yamamoto and Church (2001) show how to use suffix arrays to calculate a number of statistics that

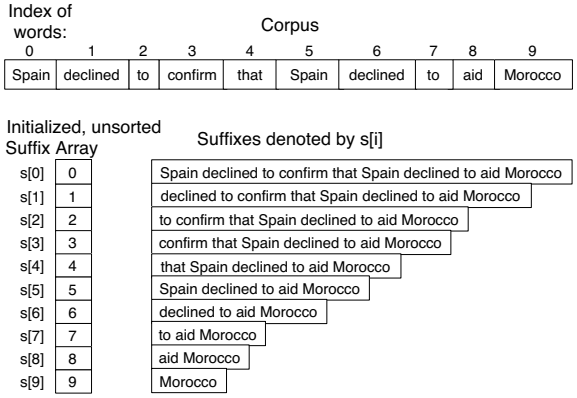

Figure 3: An initialized, unsorted suffix array for a very small corpus

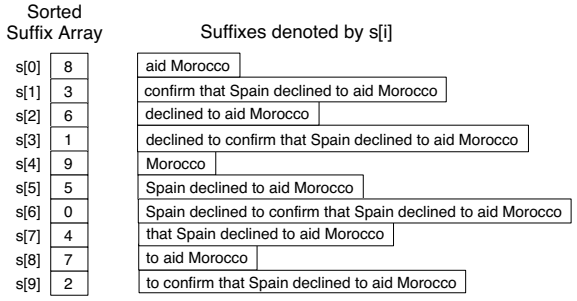

Figure 4: A sorted suffix array and its corresponding suffixes

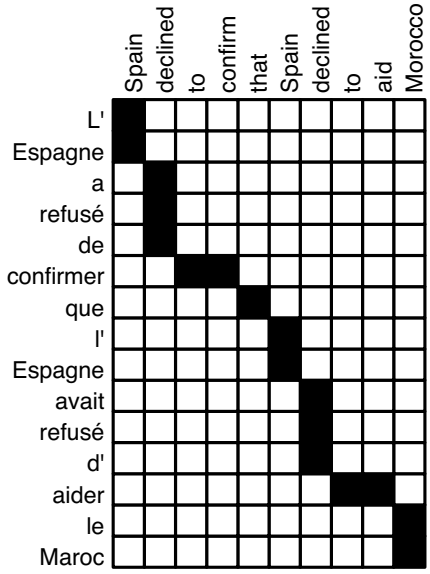

Figure 5: A word-level alignment for the sentence in the suffix array

are interesting in natural language processing applications. They demonstrate how to calculate term frequency / inverse document frequency (*tf / idf*) for all n-grams in very large corpora. Here we show how to apply suffix arrays to parallel corpora to look up the possible translations of a phrase.

4.1 Applied to parallel corpora

Our suffix array-based data structure consists of the following components:

- A sentence aligned parallel corpus such as a translation memory.
- A suffix array created for the source language portion of the corpus, and another created for the target language portion of the corpus if we are searching for phrases in both languages.
- Word-level alignments for each sentence pair in the parallel corpus.
- An array that associates word indices with sentence pairs in the parallel corpus, and their word alignments.

We could create a searchable translation memory with a parallel corpus consisting of the one sentence given in Figure 3, its French translation and the word alignment between the two sentences (given in Figure 5). We could query this searchable translation memory with the English phrase *Spain declined*. Using the suffix array given in Figure 4 we can find the indices in the source corpus where each occurrence of the phrase begins (positions 5 and 0). Then using the phrase alignment definition given in Equation 2.2 we can extract the French phrases *l'Espagne avait refusè* and *l'Espagne a refusè* as potential translations of the phrase.

In the normal case, a searchable translation memory will consist of thousands of aligned sentences. In this case we will often have multiple possible examples of each translation. We group these and display them in a limited number of contexts as shown in Figure 1. We can rank each possible translation based on its frequency (which as the same as ranking them with the translation probability defined in Equation 2.2). In the case of our one sentence searchable translation memory, the possible translations would have equal rank.

Thus we are able to use our suffix array-based data structure to find a set of possible phrase translations, and to rank them based on their probabilities. In the next section we evaluate how often a particular searchable translation memory was able to retrieve the correct translations of query phrases.

5 Evaluation

In order to evaluate our searchable translation memory we first constructed a sentence-aligned translation memory using 50,000 sentences from the German-English section of Europarl Corpus (Koehn, 2002). We selected a set of 120 German phrases to use as query terms, and retrieved all sentence pairs containing those phrases. We had two bilingual native German speakers manually align the German phrases to their English counterparts, thus creating a "gold standard" set of data for those phrases. We were able to measure the precision and recall of our automatic indexing and phrase ranking techniques against these gold standard alignments.

Precision and recall are the standard evaluation techniques for information retrieval, and are defined as follows:

- *Precision* – the ratio of phrases which we correctly retrieved to the total number of phrases

that were retrieved

- *Recall* – the ratio of phrases in gold standard which were retrieved to the total number of phrases in gold standard

For these phrases we had an average precision of 77.98%, and an average recall of 81.62%.

Since the first few phrase translations are more important to a user, we also measured how often the top ranked phrase was a correct translation. We found that the average precision for the first phrase that our system returns for this data set was 86.6%.

6 Related Work

The idea of integrating machine translation as an appropriate aid to the human translation process is not a new one. Kay (1980) described a translator's workbench, and argued that the proper place for machines in translation is as a non-obtrusive aid to human translators. The Transtype project (Foster et al., 2002) has investigated the use of statistical machine translation in particular for text prediction for human translators. Church and Hovy (1993) describes creative ways in which machine translation can be usefully applied.

Our research also relates to the evaluation of statistical models of translation. Previous work in this area has focused on the automatic evaluation of machine translation systems Bleu (Papineni et al., 2002), and on the accuracy of automatic alignments (Och and Ney, 2003). Searchable translation memories can be thought of as a task-based evaluation of statistical translation models.

7 Discussion

In this paper we investigate a useful application for machine translation technology in its current state. Rather than use statistical machine translation to perform fully automated translation, we have shown how it might instead be integrated into the human translation process by increasing the utility of translation memories. Existing translation memories only allow the reuse of previous work when whole sentences are matched in the database. Our technology allows a user to retrieve previous translations for smaller units such as phrases, without having to trawl through sentences using a concordancer. We effectively exploit the technology developed for statistical machine translation and repurpose it for another task. We have shown with the example English-German translation memory that an accuracy of nearly 90% can be achieved, which suggests the technology might be a valuable addition to a translator's workbench. Furthermore, by indexing our searchable translation memories with our memory-efficient suffix array-based method of storing phrases, we have shown a technology can be easily run within the memory available on a standard desktop computer.

Acknowledgments

The authors would like to thank Beatrice Alex and Marco Kuhlmann for their valuable assistance with creating the evaluation set.

References

Michael Barlow. 2004. Parallel concordancing and translation. In *Proceedings of ASLIB Translating and the Computer 26*.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Poossin. 1988. A statistical approach to language translation. In *12th International Conference on Computational Linguistics*.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of ACL*.

Kenneth Church and Eduard Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8.

George Foster, Philippe Langlais, Elliott Macklovitch, and Guy Lapalme. 2002. Transtype: Text prediction for translators. In *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Demonstration Description.

William Gale and Kenneth Church. 1993. A program for aligning sentences in bilingual corpora. *Compuatational Linguistics*, 19(1):75–90.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Compuatational Linguistics*, 19(1):121–142.

Martin Kay. 1980. The proper place of men and machines in language translation. Technical Report CSL-80-1, Xerox PARC. Reprinted in Machine Translation 12(1):3–23, 1997.

Kevin Knight. 1999. A statistical MT tutorial workbook. Prepared for the 1999 JHU Summer Workshop.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished Draft.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*.

Udi Manber and Gene Myers. 1990. Suffix arrays: A new method for on-line string searches. In *The First Annual ACM-SIAM Symposium on Dicrete Algorithms*, pages 319–327.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. In *Proceedings of EMNLP*.

Arturo Trujillo. 1999. Translator's workbench and translation aids. In *Translation Engines: Techniques for Machine Translation*, pages 57–85. Springer-Verlag.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *Proceedings of MT Summit 9*.

Mikio Yamamoto and Kenneth Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Compuatational Linguistics*, 27(1):1–30.