Searchable Translation Memories

Chris Callison-Burch, Colin Bannard & Josh Schroeder
Linear B Ltd.
39 B Cumberland Street
Edinburgh EH3 6RA
{chris,colin,josh}@linearb.co.uk

Abstract

In this paper we introduce a technique for creating searchable translation memories. Linear B's searchable translation memories allow a translator to type in a phrase and retrieve a ranked list of possible translations for that phrase, which is ordered based on the likelihood of the translations. The searchable translation memories use translation models similar to those used in statistical machine translation. In this paper we first describe the technical details of how the TMs are indexed and how translations are assigned probabilities, and then evaluate a searchable TM using precision and recall metrics.

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. However they have significant limitations. This paper describes a translation memory technology that aims to take greater advantage of the intellectual property of translators.

Existing translation memory systems work by trying to find a full translation unit in the database of translations that exactly or approximately matches the user's input text (Trujillo, 1999). A unit usually means a sentence or a paragraph. Most TM systems provide automatic alignment of sentence and paragraph units. Although it is possible to perform hand alignment of sub-sentential units, it is a time-consuming process. Only allowing the matching of sentences means very limited use is made of the translation archive. A translator will frequently be using phrases, words or other subsentential strings that s/he has translated many times before. However, unless there is a whole unit in the database that matches this string, the system will not be able to retrieve translations for these from the database.

Having a translation database that can be freely searched, and is able to return previous translations of a user's input where that input (or a part of that input) matches only a part of a previous segment promises to greatly increase the usefulness of the archive to a translator. This paper describes a system that offers precisely that facility. This tool is the result of research into exploiting translation memories for the building of machine translation systems.

Statistical machine translation (Brown et al., 1993) is a data-driven approach to machine translation. Rather than relying on the more conventional approach of having a team of linguists and lexicographers laboriously hand-craft rules, statistical machine translation systems are created by automatically analyzing a *parallel corpus*. A parallel corpus is a collection of example translations which have been aligned on the sentence level, such as a translation memory. By examining the co-occurrence of the words in one language paired with words in another language a bilingual dictionary is induced. The grammatical structure of the languages is very

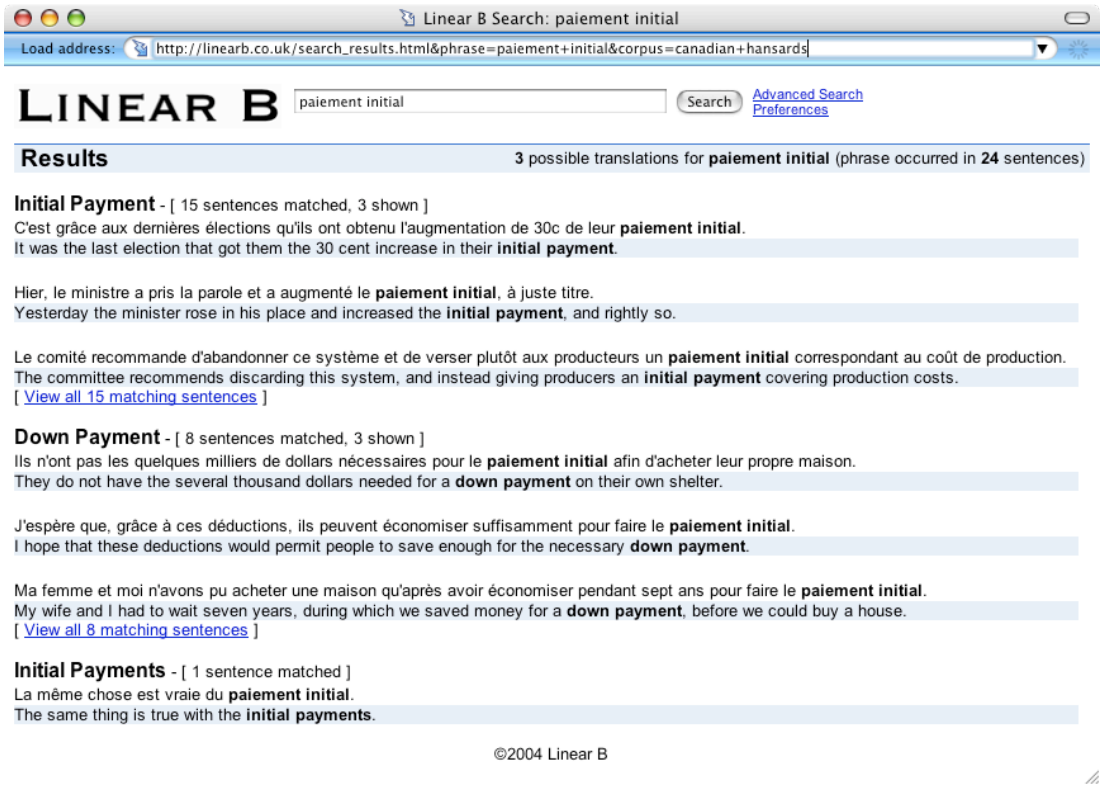


Figure 1: Search results for the French phrase “paiement initial”

roughly approximated by examining the relative ordering of the words.

Statistical machine translation has recently made very significant leaps forward in quality. Even in view of these advances in quality, statistical machine translation has not yet achieved the holy grail of machine translation: it has not yet proven to be an useful tool for human translators. Linear B is a commercial provider of statistical machine translation systems. This paper describes Linear B’s first foray into building software aides for human translators. Rather than using the statistical translation models to produce translations through fully-automatic machine translation, we instead use the models in order to allow Google-style searching of translation memories.

Figure 1 shows example results of querying a searchable translation memory built from the proceedings of the Canadian Parliament. The user has typed the search phrase *paiement initial*, and similar to a concordancer, the Linear B system has returned a list of sentences that the phrase occurs in. However, unlike a concordancer, the searchable translation memory picks out which phrases constitute the likely translations (*initial payment*, *down payment*, and *initial payments*), and ranks them according to their probability.

The novelty of our method is that it suggests the translation that the user is looking for, and highlights the relevant section of the retrieved sentences, rather than just returning a jumble of matching sentences. The technical feat in the method is the indexing of a translation memory such that the correspondence in translated words and phrases is discovered. For this our searchable translation memories rely on the data-driven methods used in statistical machine translation. These methods have the advantage of being language independent, meaning we can build a searchable translation memory from a translation archive that contains text in any languages.

The remainder of the paper is as follows: Section 2 describes how we index a searchable translation memory and rank the phrases that we retrieve using alignments produced with statistical models of

translation. In section 3 we use precision and recall (which are standard evaluation measures for information retrieval systems) to evaluate a searchable translation memory built using text from the European Parliament. Section 4 shows how the process can be extended so that even if a query phrase is not found in the TM, a translation may still be retrieved. Section 5 discusses the potential application of statistical translation models to the human translation process more generally.

2 Indexing a Translation Memory

The foundation of our searchable translation memories is our technology for automatically aligning phrases across the two languages. This is a difficult task because in general the only units that are aligned in a translation memory are sentence pairs (though occasionally technical terminology is also aligned). Having the computer discover which words and phrases correspond without being able to rely on a dictionary is not a simple task.

In order to discover these correspondences, we are able to draw on more than a decade's worth of research in the field of statistical machine translation. Statistical machine translation provides mechanisms for learning word-level alignments from the larger sentence-aligned units in parallel corpus. From these word-to-word alignments, we can extrapolate phrase-to-phrase correspondences. Once we have enumerated all phrase-to-phrase correspondences in a parallel corpus, it is relatively straightforward to build an index which allows us to retrieve a set of possible translations for a certain phrase, show the original contexts which they appeared in, and rank the translations based on their likelihood of being correct.

Sections 2.1 and 2.2 give a brief overview of statistical machine translation, along with references to more information. Readers who are not mathematically inclined may freely skip those sections, but are encouraged to examine Figures 3 and 2, which show how phrases are extracted and indexed for retrieval.

2.1 Word Alignments

The goal of statistical machine translation is to be able to determine how likely a translation is given a particular source sentence. This likelihood allows us to construct a set of candidate translations, and to rank them based on their probabilities. Brown et al. (1993) formulated translation as essentially 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 possible word-level alignments, $\mathbf{a}$, between the sentences:

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

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. Figure 2 illustrates a probable word-level alignment between a sentence pair in the Canadian Hansard bilingual corpus.

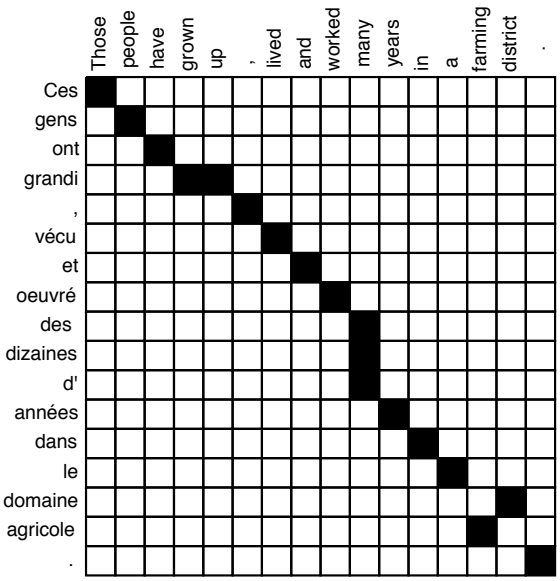
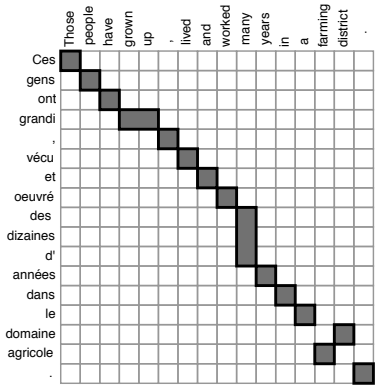


Figure 2: A word-level alignment for a sentence pair that occurs in our training data

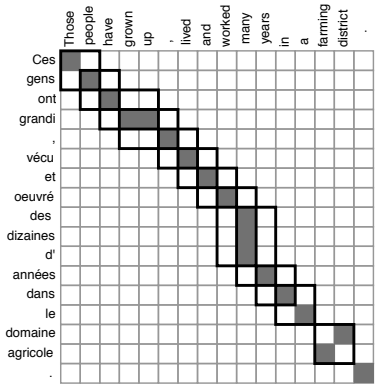
2.2 Phrase Alignments

Whereas the original formulation of statistical machine translation was word-based, more contemporary approaches have expanded to using phrases. Phrase-based statistical machine translation uses larger segments of human translated text. In general, the probability of an English phrase $\bar{e}$ translating as a French phrase $\bar{f}$ is calculated as the number of times



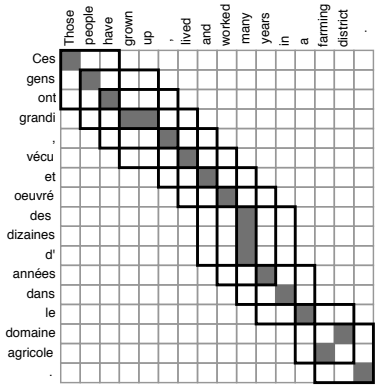
Phrases extracted on iteration 1: (Ces : Those), (gens : people), (ont : have), (grandi : grown up), (, : ,), (vécu : lived), (et : and), (oeuvré : worked), (des dizaines d' : many), (annés : years), (dans : in), (le : a), (domaine : district), (agricule : farming), (. : .)

Notice that one word can translate as a phrase, such as 'grandi' → 'grown up'. Incrementally larger phrases are created by incorporating adjacent phrases words and phrases.



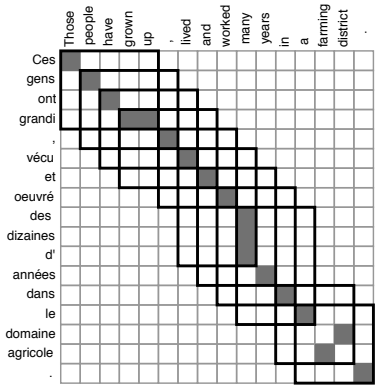
Iteration 2: (Ces gens : Those people), (gens ont : people have), (ont grandi : have grown up), (grandi , : grown up ,), (, vécu : , lived), (vécu et : lived and), (et oeuvré : and worked), (oeuvré des dizaines d' : worked many), (des dizaines d' annés, many years), (annés dans : years in), (dans le : in a), (domaine agricule : farming district)

Notice that 'a farming' does not have a translation since the phrase that it spans is not contiguous.



Iteration 3: (Ces gens ont : Those people have), (gens ont grandi : people have grown up), (ont grandi , : have grown up ,), (grandi , vécu : grown up , lived), (, vécu et : , lived and), (vécu et oeuvré : lived and worked), (et oeuvré des dizaines d' : and worked many), (oeuvré des dizaines d' annés : worked many years), (des dizaines d' annés dans : many years in), (annés dans le : years in a), (le domaine agricule : a farming district), (domaine agricule . : farming district .)

Notice that 'a farming district' now has a translation since the phrase 'farming district' is adjacent to the phrase 'a'.



Iteration 4: (Ces gens ont grandi : Those people have grown up), (gens ont grandi , : people have grown up ,), (ont grandi , vécu : have grown up , lived), (grandi , vécu et : grown up , lived and), (, vécu et oeuvré : , lived and worked), (vécu et oeuvré des dizaines d' : lived and worked many), (et oeuvré des dizaines d' annés : and worked many years), (oeuvré des dizaines d' annés dans : worked many years in), (des dizaines d' annés dans le : many years in a), (dans le domaine agricule : in a farming district), (le domaine agricule . : a farming district .)

Figure 3: Extracting incrementally larger phrases from a word alignment

that the English phrase was aligned with the French phrase in the training corpus, divided by the total number of times that the French phrase occurred:

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

The trick is how to go about extracting the counts for phrase alignments from a training corpus.

Many methods for calculating phrase alignments use word-level alignments as a starting point.[1] There are various heuristics for extracting phrase alignments from word alignments, some are described in Koehn (2003), Tillmann (2003), and Vogel et al. (2003).

Figure 3 gives a graphical illustration of the method of extracting incrementally larger phrases[2] from word alignments described in Och and Ney (2003). Counts are collected over phrases extracted from word alignments of all sentence pairs in the training corpus. These counts are then used to calculate phrasal translation probabilities.

2.3 Indexed Phrases

Linear B's searchable translation memories are built from phrases extracted using the method illustrated in Figure 3. We create an index containing each phrase, its possible translations, and a link back to the original sentences. A record in our index stores the following information:

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

After we have built 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 described in Section 2.2.

We can then build suitable interfaces that will allow a translator to search the database via a simple search window, or through some other mechanism. The results of the search can be presented as a list of possible translations with the most probable translation first. The context sentences from which the translation was extracted may also optionally be displayed, as shown in Figure 1.

3 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 = number of phrases which we correctly retrieved / total number of phrases that we retrieved
RECALL = number of phrases in gold standard which we retrieved / total number of phrases in gold standard

Figure 4 gives the a sample of results for the experiment. The first column shows the German phrase, the second column shows the set of phrases that our automatic method extracted, ant the third column shows the gold standard translations. For these sentences we had an average precision of 77.98%, and an average recall of 81.62%.

Since the first few phrases are more important to the user's experience, 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%.

4 Extensions

These evaluation results are impressive. They suggest that our method may allow a user to efficiently

[1] 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.

[2] Note that the 'phrases' in phrase-based translation are not congruous with the traditional notion of syntactic constituents; they might be more aptly described as 'substrings' or 'blocks'.

German phrase	Automatically extracted translations	Gold standard
absolut klares	crystal clear	crystal clear
am Arbeitsplatz	at work, workplace, the work	at work, at the workplace, in the work sphere
besondere Aufmerksamkeit	particular attention, special attention, pay particular attention, great care, specific attention, attention	particular attention, special attention, great care, greater emphasis, specific attention
der Kohlendioxid-emissionen	carbon dioxide	of carbon dioxide emissions
Deutlichkeit	clarity, clearly, clear, quite unequivocally, crystal clear, make, it	clarity, clearly, clear
durchschnittlich	average, an average, on average	on average, average
eine Million	one million, a million	one million, a million
eingedämmt	checked, under control, is, curbed	checked, slow down, under control, curb, curbed, limit
erste Halbjahr	first half	first half of the year
ersten Hälfte	first half	first half
früher oder später	sooner or later	sooner or later, at some point, eventually
ganz klar	quite clear, clear, very clear, quite clearly, clearly, very clearly, very sure, a clear, let, quite explicitly, crystal clear, no uncertain, very well, is clear, very unclear, absolutely clear, particularly clear, perfectly clear, quite openly	quite clear, clear, very clear, quite clearly, clearly, obviously, very clearly, very sure, quite explicitly, crystal clear, obvious, no uncertain, very well, very unclear, no mistake, absolutely clear, particularly clear, perfectly clear, in great detail, quite openly
große Sorgfalt	great care	great care
große Vorsicht	great care, a careful, very careful	great care, very careful, a careful approach
Großkapitals	big, capitalism, big business	big investors, big capital, capitalism
grüne Licht	green light	green light
grünes Licht	green light, ahead	green light, go-ahead, the formal go-ahead, approval
im Durchschnitt	on average	an average, on average
im Zaume	under control	under control
Lärmbelastung	noise, noise pollution, noise exposure	noise pollution, noise, noise exposure
sehr darauf	great care, extreme caution	extreme caution
sozialistischen Partei	socialist party	socialist party
Streitkräfte	forces, armed, defence, military force, military forces	forces, military forces, defence, military force, armed forces
Truppe	force, military force, will	force, military force
unter Kontrolle	under control, in check	under control, in check

Figure 4: A sample of the evaluation data used to produce precision and recall results

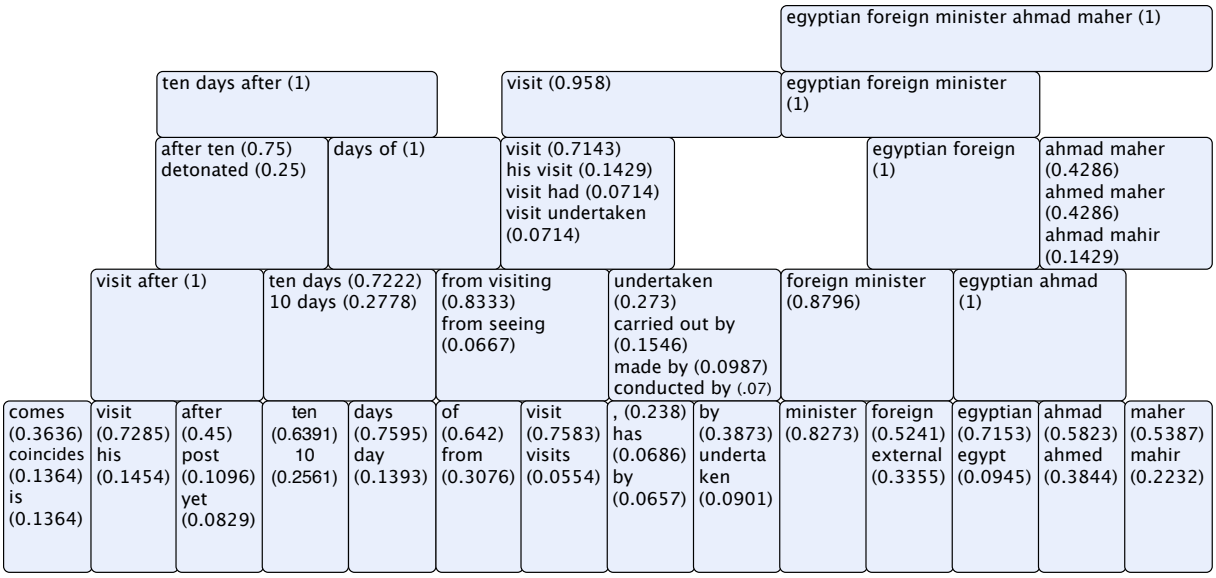


Figure 5: A chart of the various sub-sentential segments that were retrieved for an Arabic sentence, along with their associated probabilities

retrieve phrases from a translation memory without have to trawl through sentences by hand using a concordance. One additional, hitherto undiscussed, advantage of our searchable translation memories over concordances is this: our searchable translation memories are not limited to phrases which appear in the corpus. Even given a large database of translations, a user will want to locate translation for phrases which are not contained in their entirety. Because our system borrows from techniques in data-driven machine translation, we can intelligently construct translations of unseen phrases from smaller seen units.

This can be done as follows. Given an input of any length, we break that input into all possible substrings, and look up all translations that were aligned with each substring in the training corpus. We then search through all possible combinations of these substrings in order to find the best translation of source input. We do this using our translation probabilities and an additional probability for the fluency of the translation.

As with the phrase retrieval we are able to offer not just a single candidate translation of the input, but rather a list of possibilities ranked according to their probability. Figure 5 shows an example chart of possible translations and the probabilities that are used to choose between them. From this our software is able to reconstruct a translation on the entire sentence, or of pieces of it, whether or not they were found as phrases in the translation memory.

5 Discussion

Linear B is a commercial provider of statistical machine translation systems. We believe that there are many features of our data-driven technology that make it a more appropriate tool for human translation than previous machine translation technologies were. Because the output is derived from large segments of human translated text it is higher in quality. Because a system can be trained from any collection of example translations, it is more easily customizable to specialised domains. And because our technology continues to learn from translators' output, it moves closer to the needs of our customers as they use it. We believe that these factors make our systems more appropriate for use as a human translation

aid.

This paper described Linear B's first foray into building software aides for human translators. Rather than using our technology 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. This is based on an intelligent technology that automatically finds alignments of words and phrases. An English translation memory of 1.5 million sentences produced 10 million possible translation units for retrieval, increasing the usefulness of the archive almost sevenfold. We have shown that for an example English-German translation memory our system returns a perfect translation of the input phrase nearly nine times out of ten.

This paper also described an extension of this method that allows us to offer a translation of a user's input even when that input cannot be found in the archive. This is a synthesis of the TM and MT technologies. By retrieving translations of parts of the input and intelligently recombining them we can return a new translation of the input, even if it is not found in whole in the archive. By applying technology developed for statistical machine translation to translation memories, the tools described here are aimed at maximizing the utility of a translator's intellectual resources. We believe that these applications hail the arrival of a new generation of intelligent translation memories.

References

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.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Version 2 of the corpus.

Philipp Koehn. 2003. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. User Manual and Description.

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.

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

Arturo Trujillo. 1999. *Translation Engines: Techniques for Machine Translation*. 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*.