Improved Statistical Translation Through Editing

Chris Callison-Burch, Colin Bannard *and* *Josh Schroeder*

School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
{chris,colin}@linearb.co.uk

Linear B Ltd.
Technology Transfer Centre
King's Buildings
Edinburgh EH9 3JL
josh@linearb.co.uk

Abstract. In this paper we introduce Linear B's statistical machine translation system. We describe how Linear B's phrase-based translation models are learned from a parallel corpus, and show how the quality of the translations produced by our system can be improved over time through editing. There are two levels at which our translations can be edited. The first is through a simple correction of the text that is produced by our system. The second is through a mechanism which allows an advanced user to examine the sentences that a particular translation was learned from. The learning process can be improved by correcting which phrases in the sentence should be considered translations of each other.

.

1. Introduction

Statistical machine translation was first proposed in Brown et al (1988). Since statistical machine translation systems are created by automatically analyzing a corpus of example translations they have a number of advantages over systems that are built using more traditional approaches to MT:

- They make few linguistic assumptions and can therefore be applied to nearly any language pair, given a sufficiently large corpus.
- They can be developed in a matter of weeks or days, whereas systems that are hand-crafted by linguists and lexicographers can take years.
- They can be improved with little additional effort as more data becomes available.

More recent advances in phrase-based approaches to statistical translation (Koehn et al (2003), Marcu and Wong (2002), Och et al (1999)) have led to a dramatic increase in the quality of the translation systems. Phrase-based translation systems produce higher-quality translation since they use longer segments of human translated text. Using longer segments of human translated text reduces problems associated with literal word-for-word translations. For example, multi-word expressions such as idioms are better translated.

Linear B is a commercial provider of statistical machine translation systems. This paper describes Linear B's advances to phrase-based machine translation that allow translation quality to be improved through editing translations that are produced by our system. There are two levels at which our translations can be edited:

- The first is through a simple correction of the text that is produced by our system. Our system improves by dynamically learning the correct translations of new phrases. These new phrases are extracted from the corrected sentence pair using the existing translation models, and can be used immediately for subsequent translations.
- The second is through a mechanism that allows an advanced user to inspect which phrases the system's translation was composed from. If a particular phrase was mistranslated, the user can examine the set of sentence pairs that a particular translation was learned from, and make corrections.

These features mean that our system is capable of improving with use and adapting to be more appropriate for a new domain. This has two main implications: our systems get better as our customers use them, and our systems have the potential to be trained using example translations from one domain (such as government documents, which have abundant translations) and gradually adapted to a new domain.

The remainder of the paper is as follows: Section 2 describes how our phrase-based models of translation are learned from archived translations, and gives example output produced by a system trained on data from the Canadian parliament. Section 3 shows how our system dynamically integrates edited output by extracting the translations of new phrases, and weighting the corrected translations more heavily than existing translations in the model. Section 4 described the advanced editing technique that allows a user to inspect the sentence pairs which a faulty translation was learned from, and correct the statistical models by explicitly showing the system which phrases ought to be learned from those sentence pairs instead.

2. Phrase-based Statistical Translation

The goal of statistical machine translation is to be able to choose that English sentence, **e**, that is the most probable translation of a given sentence, **f**, in a foreign language. Rather than choosing **e*** that directly maximizes the conditional probability p(**e**|**f**), Bayes' rule is generally applied:

$$\mathbf{e}^* = \text{argmax}_e\, p(\mathbf{e})\, p(\mathbf{f}|\mathbf{e})$$

The effect of applying Bayes' rule is to divide the task into estimating two probabilities: a language model probability p(**e**) which can be estimated using a monolingual corpus, and a translation model probability p(**f**|**e**) which is estimated using a bilingual, sentence-aligned corpus. Here we examine different ways of calculating the translation model probability.

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

Figure 1. A word-level alignment for a sentence pair that occurs in our training data

2.1 Word Alignments

Brown et al (1993) define a series of translation models, which are commonly referred to as IBM Models 1 to 5. The IBM Models formulate translation essentially as a word-level operation. The probability that a foreign sentence is the translation of an English sentence is calculating by summing over the probabilities of all possible word-level alignments, **a**, between the sentences:

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

Thus they 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 1 illustrates a probable word-level alignment between a sentence pair in the Canadian Hansard bilingual corpus.

Brown et al formulate alignment probability p(**f,a**|**e**) in terms of *distortion*, *fertility*, and *spurious word* probabilities in addition to word-for-word translation probabilities. The act of translation in the Brown et al approach is one of *string rewriting*. In string rewriting each word in a source sentence is replaced by zero or more words in the target language. Then, a number of ``spurious'' target words might be inserted with no direct connection to the original source words. Finally the words are then reordered in some fashion to form the translation.

Problems with the Brown et al approach to translation include:

- It doesn't have a direct way of translating phrases; instead fertility probabilities are used to replicate words and translate them individually.
- Using small units such as words means that a lot of word reordering has to happen. But the distortion probability is a poor explanation of word order.

2.2 Phrase Alignments

Phrase-based translation, by contrast, uses larger segments of human translated text. Phrase-based translation does away with fertility and spurious word probabilities. While it does have some notion of distortion, this is less pertinent since local reorderings such as adjective-noun alternation can be easily captured in phrases. The main part of phrase-based translation models is the estimation of phrasal translation probabilities.

In general, the probability of an English phrase *e* translating as a French *f* is calculated as the number of times 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:

$$p(f|e) = count(f,e) / \sum_f count(f,e)$$

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).

The online version of this paper 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.

The act of translating with phrase-based translation model involves breaking an input sentence into all possible substrings, looking up all translations that were aligned with each substring in the training corpus, and then searching through all possible translations to find the best translation of source sentence.

2.3 Example Translation

This section gives an example translation which was produced by our system when it was trained on a collection of example translations from the Canadian Parliament. For the source passage

> L'honorable Leonard J. Gustafson: Honorables sénateurs, tandis que la guerre en Irak entre dans sa troisième semaine, nous ne devons pas oublier qu'il faut prendre des mesures pour éviter une crise humanitaire dans la population civile. À cet égard, il y a de nombreux domaines dans lesquels les Canadiens doivent faire preuve de leadership.
>
> Maintenantque la guerre fait rage en Irak et que des pénuries de produits alimentaires et de fournitures médicales commencent à se produire, les pays qui ont accès à des ressources ont la responsabilité d'essayer de minimiser les effets négatifs du conflit sur la population irakinne. Je crois personnellement que le Canada devrait jouer un plus grand rôle à cet égard.
>
> Au Canada, nous disposons en abondance de blé et d'autres produits alimentaires. Nous pouvons également fournir des articles médicaux aux citoyens de l'Irak. Le Canada a une fière réputation pour ce qui estd'offrir une aide humanitaire aux gens quand ils en ont besoin.

our system produced the translation

> The Honourable Leonard J. Gustafson: Honourable senators, while the war in Iraq extending into a third week, the minister should not forget the one we must take some steps to prevent a crisis humanitarian face in the civilian populations. In this regard, there are a number of areas in which the Canadians must concentrate to show leadership.
>
> That the war is raging in Iraq and that a shortage of food and medical supplies are beginning to take place, those countries that have access to the resources are the responsibility for doing try to minimize the negative effects on workers on the people Irakienne. I personally believe that Canada should play a more significant role in this regard.
>
> In Canada, we have in abundance of wheat, as have other food products. We can also provision of medical articles to the people on the other Iraq. Canada has a proud record in so far, as would be to give humanitarian assistance to people when they need it.

As a reference here is how a human translator rendered the passage

[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 the traditional notion of syntactic constituents; instead they might be more aptly described as `substrings' or `blocks'.

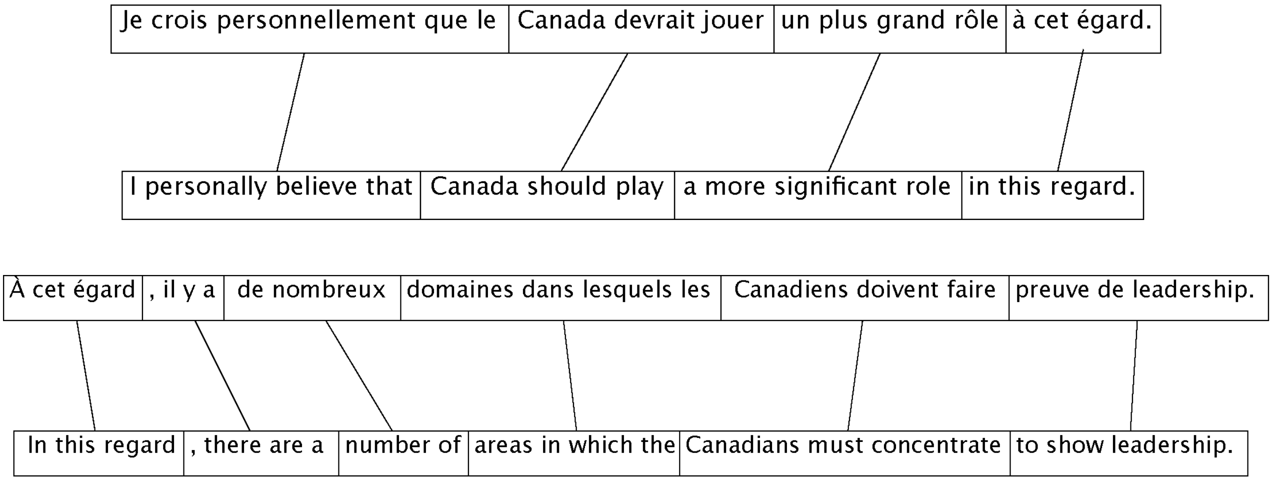


Figure 2. An example of the phrases that were used to translate two sentences

> Hon. Leonard J. Gustafson: Honourable senators, as the war in Iraq enters its third week, the need to take measures to avoid a humanitarian crisis among Iraq's civilian population should not be forgotten. In this regard, there may be avenues where Canadians must provide leadership.
>
> As the war wages in Iraq and shortages of foodstuffs and medical supplies start to occur, countries with access to resources have a responsibility to help minimize the negative impact of the conflict on the country's population. In this regard, it is my belief that Canada should play a greater role.
>
> In Canada, we have access to plentiful supplies of wheat and other foodstuffs. Canada can also provide medical supplies to the citizens of Iraq. Canada has a proud history in providing humanitarian assistance to people in times of need.

Figure 2 shows which phrases were selected by our system when deciding how to translate two of the sentences from the passage.

3. Simple Editing

Because our translations are learned from example translations we can exploit edited system output to improve the quality of subsequent translations. The most straightforward way of improving the translation quality would be to add the edited translations (along with the source sentences that they were produced from) to the training corpus, and re-train the system on the augmented set of data. However, training the system can take a long time -- it took longer than a week to train the system using the parallel corpus containing roughly 25 million words of Canadian Parliament text. Since it takes so long to re-train a system, this method could only be done periodically, and translation quality would not immediately benefit from a user correcting the system's translations.

Instead we have developed what we term an ``alignment server''. The alignment server is a variant of the software that trains the translation models. It keeps the parameters of a translation model in memory, and is able to create a word alignment on-the-fly from a new sentence pair.

For example, if our system had translated the French sentence

> *Durant l'ère glacière, les changements climatiques obligèrent le règne animalèa migrer vers des contrées plus accueillantes.*

into English as

> *During the era refrigerator, the climatic changes oblige the the reign animal to migrate towards more accessible regions.*

A user would easily be able to edit the translation to make it better:

> *During the ice age, climatic changes caused the animal kingdom to migrate towards more accessible regions.*

Our alignment server then produces a word-level alignment such as the one depicted in Figure 3. From this we extract a set of new phrases and their translations. We extract *the ice age* as the correct translation of the phrase *l'ère glacière*, and *animal kingdom* as the correct translation of *règne animale*. These would be added to the database of phrasal translations that the decoder draws upon, and would prevent the incorrect translation of *l'ère glacière* as *the era refrigerator* or *règne animale* as *reign animal* in the future.

Since our alignment server produces new word alignments dynamically as text is edited, we are able to update our database in real time. In this way we are able to take advantage of new phrases almost immediately. A problem occurs when adding new translations for phrases which already exist in the database. The problem is that the there are usually many more occurrences of the phrase in the existing training data.

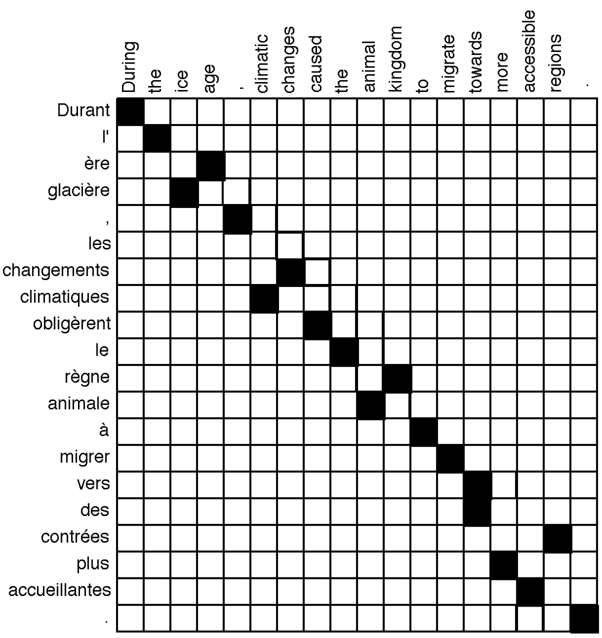


Figure 3. Word alignment produced by the alignment server for an edited translation

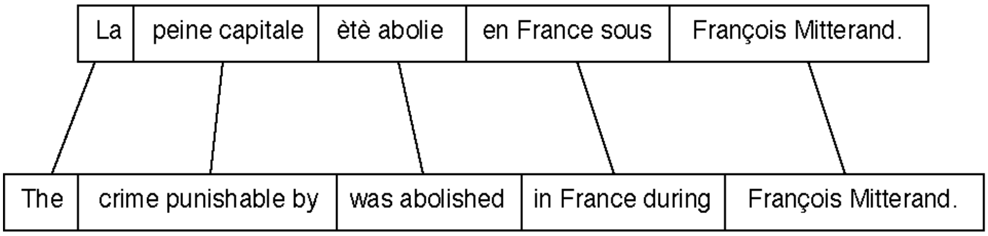


Figure 4. Inspecting the phrases that were used to generate a translation to discover misaligned phrases.

We would like the edited text to play an important role in shaping what translations are produced, since it will reflect how our customers would like the text to be translated. However, since the amount of edited text will be dwarfed by the amount of text in the training data, the translations will likely still look very similar to the training data. To overcome this, we employ a weighting scheme in calculating the probabilities.

$$p(f|e) = \lambda_1 \text{count}_{C1}(f,e) + \lambda_2 \text{count}_{C2}(f,e) \ / \ \lambda_1 \sum_f \text{count}_{C1}(f,e) + \lambda_2 \sum_f \text{count}_{C2}(f,e)$$

The λ_1 and λ_2 co-efficients allow us to scale the contribution of phrase alignments extracted from the newly created corpus (C2) of edited translations, and downweight the counts of phrases from the original training corpus (C1). This allows us to create a system from existing public domain data, such as the translations of government proceedings, and gradually adapt it to a new domain, while placing more weight on the newly created data.

4. Advanced Editing

As an alternative to adding new phrasal translations to the database, we have also implemented a feature which allows advanced users to identify and correct the underlying cause of a mistranslation. Often a poor translation will arise because phrases were incorrectly extracted from the training data due to misalignment between words in the sentence pairs. We allow a user to inspect all of the sentence pairs in the training data that the incorrect phrase was learned from.

For example, the French sentence

> *La peine capitale ètè abolie en france sous François Mitterand.*

was translated by our system as

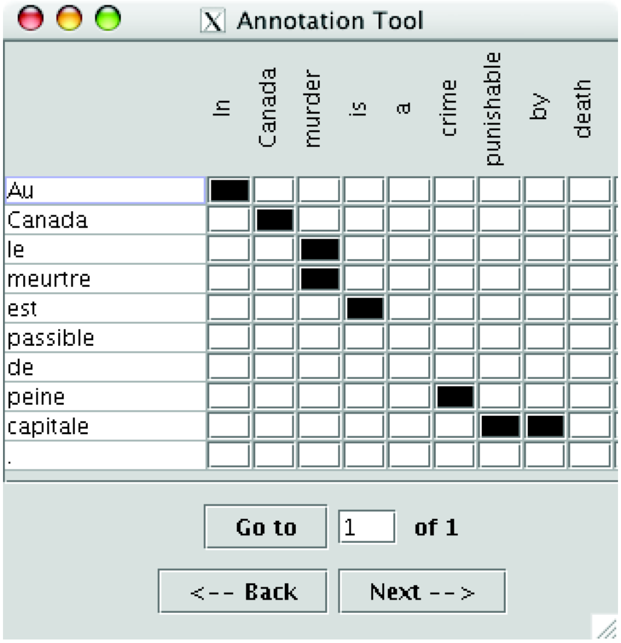
The crime punishable by was abolished in France during François Mitterand.

Simple editing could be performed to change the translation to

The death penalty was abolished in France under François Mitterand.

We can also inspect the phrases that were selected in order to produce this translation (they are shown in Figure 4). By identifying the phrase pairs which are poor translations, such as *peine capitale* being translated as *crime punishable by*, we can query the database and inspect which sentence pairs in our training data the phrasal translation was learned from.

We can then retrieve those sentence pairs and manually correct their automatically generated word alignments using the software tool that is shown in Figure 5. As a user updates the word-level alignment, another window displays the changes to the phrases that are extracted from the sentence pair. Figure 6 shows how the problematic word alignment can be improved slightly, by eliminating partial translations. After pressing the *Save* button the counts for three incorrect phrasal translations would be decremented in the database, and the count for the correct translation would be incremented.

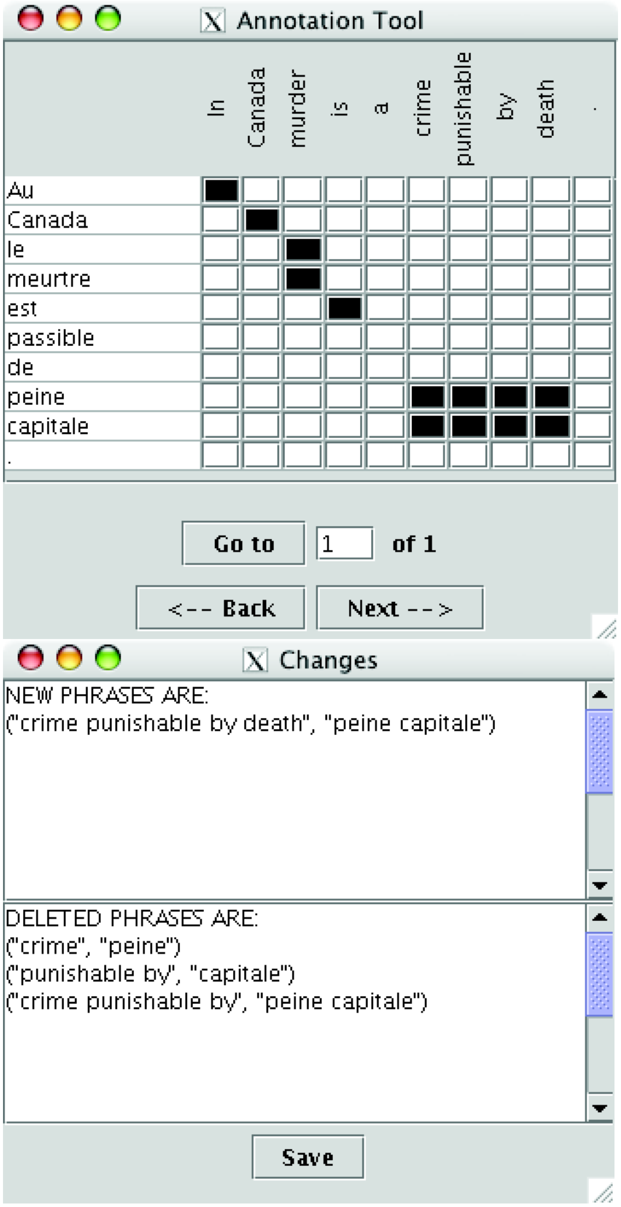


	In	Canada	murder	is	a	crime	punishable	by	death
Au	■								
Canada		■							
le			■						
meurtre			■						
est				■					
passible									
de									
peine						■			
capitale							■	■	
.									

Figure 5. A software tool which allows an advanced user to correct misaligned sentence pairs from the training data

By allowing the user to inspect exactly which sentences any phrasal translations were learned from we hope to demystify statistical machine translation. A user can find out why our system translated a sentence in a certain way, and change the behavior of the system by altering its underlying representation.

Moreover, Callison-Burch et al (2004) shows that training data that has been manually aligned on the word-level can be used to significantly improve the alignment accuracy and the translation quality of statistical machine translation. Thus, by incorporating the manual modifications of our training data into the next round of training, we hope to improve the word alignments of the rest of the sentence pairs as well.



	In	Canada	murder	is	a	crime	punishable	by	death	.
Au	■									
Canada		■								
le			■							
meurtre			■							
est				■						
passible										
de										
peine						■	■	■	■	
capitale						■	■	■	■	
.										

NEW PHRASES ARE:
("crime punishable by death", "peine capitale")

DELETED PHRASES ARE:
("crime", "peine")
("punishable by", "capitale")
("crime punishable by", "peine capitale")

Figure 6. Sample figure and caption

5. Discussion

There are a number of limitations that make non-statistical machine translation unsuitable for the human translation process. Specifically:

- It is difficult to adapt machine translation to new domains. Fully adapting a system to a new domain is essentially not possible without having a staff of lexicographers overhaul the system at great cost.
- Those tools which some systems provide building a user dictionary are fairly crude, and often do not adequately allow the user to change the behavior of the system.

This lack of customizability acts as an obstacle to machine translation gaining use as an aid to human translation. Much translation work is extremely specialized, and a single system cannot meet the needs of translators across the range of domains.

Statistical machine translation overcomes this to some extent. Systems can be built using the existing translation archives of an individual or an organisation, and are thus customized to the particular domain that they are working in.

Linear B's system is specifically designed to bring out the elements of customizability and adaptability that are inherent in data-driven approaches. Our translation software can adapt to to the translator as it is used through a simple correction of the system's output. Further, it makes statistical machine translation transparent by allowing users to inspect how translations were learned, and to change the behavior of the system by altering the consequent representations.

We hope that these advances will make machine translation more useful in the human translation process.

References

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, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. Paper Draft.

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

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.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999.Improved alignment models for statistical machine translation. In Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora.

ChristophTillmann. 2003. A projection extension algorithm for statistical machine translation. In Proceedings of EMNLP.