A Look at English-Inuktitut Word Alignment

Fabrizio Gotti, Alexandre Patry, Guihong Cao, Philippe Langlais
RALI
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Succursale Centre-Ville
H3C 3J7 Montréal, Canada
http://rali.iro.umontreal.ca

Abstract

Statistical Machine Translation (SMT) as well as other bilingual applications strongly rely on multilingual corpora aligned at the word level. Efficient alignment techniques have been proposed but are mainly evaluated on pairs of languages where the notion of word is mostly clear. We concentrated our efforts on the English-Inuktitut word alignment task and present two approaches we implemented and combinations of both. We discuss our approaches in the light of the a shared task proposed within the *Workshop on Parallel Texts* (WPT) of the ACL 2005 conference.

1 Introduction

Statistical Machine Translation (SMT) systems use observations on a *parallel corpus* (a set of documents translated in many languages) to tune their parameters. As a preprocessing step, many of them need the parallel corpus to be aligned at the word level (Galley et al., 2004; Och and Ney, 2004; Venugopal et al., 2003; Koehn et al., 2003) therefore the translation relations between words must be identified. Some studies have shown that word alignment is not necessary (*e.g.* (Zhao and Vogel, 2005)), but the performance of these systems is not yet fully convincing.

Several efficient word alignment techniques have been proposed (Brown et al., 1993; Och and Ney, 2003), but they have been mostly evaluated on languages where each word conveys one concept. Inuktitut fails to fulfill this criterion because many concepts can be combined into a single word. In this paper, we explore two different approaches to align English and Inuktitut documents at the word level.

Two months ago, we participated in a very intriguing shared-task proposed within the *Workshop on Parallel Texts* (WPT) of the ACL 2005 conference[1]: the alignment at the word level of an English-Inuktitut parallel document (Martin et al., 2005). As is usually the case in such exercises, a lot was done within a very short amount of time, and we did not find time at all to carefully analyze what we tried. After the deadline, we conducted experiments that helped us to gain some insight into the task as well as improving our aligners.

In the following section, we expose the task and the challenges we address. In Sections 3 and 4, we present two different systems we devised to tackle the problem: a first one which sees the word alignment task as a sentence alignment task and a second one which tries to associate an English word to a substring of an Inuktitut word. The different approaches are evaluated in Section 5. Our best system, which is a combination of both approaches achieves an error rate of 32% on a section of the Nunavut Hansard. In Section 6, we discuss our experiments in the light of the systems which participated to the WPT'05 shared-task. Finally, we conclude this work in Section 7.

[1] See www.statmt.org/wpt05/

2 Task Definition

2.1 A Quick Look at Inuktitut

Martin et al. (2003) present a short description of the Inuktitut language. Inuktitut is spoken by Inuit people living in Canada, most of them in the North Eastern part of the country. This language has its own syllabic script, but a romanised equivalent exists and is standardised.

It is a very agglutinative language, with a rich morphology. Typically, a series of *morphemes* (basic units of meaning) are all suffixed to a root word, modifying it in the process. Often, this leads to a relatively long word, corresponding to many English words and concepts. Furthermore, during this concatenation, orthographic changes are very often made to the original morphemes.

An example of a word alignment excerpted from the development corpus provided by the organisers of WPT'05 is shown in Figure 1. It is interesting to note that the word *uqausiqakainnarumajunga* corresponds to six English words (and a complete clause).

Inuktitut	pijjutigillugu$_1$ innatuqait$_2$ amma$_3$ makkuttut$_4$ uqausiqakainnarumajunga$_5$
English	[In regards to]$_1$ [elders]$_2$ [and]$_3$ [youth]$_4$ [I want to make general comments]$_5$

Figure 1: An Inuktitut sentence and its English translation. Identical indices show corresponding words.

2.2 Aligning Words

Aligning words consists in identifying words that are in translation relation in a text. Usually, this involves the use of a bilingual lexicon built by counting word co-occurrences in a training corpus. But because Inuktitut is an highly agglutinative language, many words in our Inuktitut corpora appear only once (see Table 1), which makes it harder to compute statistics on them.

3 Word Alignment as a Sentence Alignment Task

A fast inspection of our material reveals that in most of the cases, the word alignment of two documents is monotonic and involves a sequence of `1-n` pairs (1 Inuktitut word is aligned to n English ones). Knowing that many sentence alignment techniques strongly rely on the monotonic nature of the inherent alignment, we suggest applying such a strategy to the word alignment task.

The sentence aligner we relied on is an in-house program called JAPA.[2] It was one of the most accurate alignment program within the Arcade evaluation campaign (Langlais et al., 1998). In a few words, JAPA begins by defining a search space in which to search for parallel sentences. This search space can be defined by a diagonal beam or by a beam that is anchored on cognates. Once the search space is set, a dynamic programming algorithm is used to find the most probable alignment following a cost that is a linear combination of the length criterion described in Gale and Church (1993) and the score based on cognates described in Simard et al. (1992). To transform the word alignment problem into a sentence alignment one, we only have to consider single sentences as documents and tokens as sentences (we define a token as a sequence of characters delimited by white space).

4 NUKTI: Word and Substring Alignment

Although using a sentence aligner is a promising approach, it does not benefit from associations between English words and Inuktitut morphemes, even though this kind of information may potentially yield better results. Indeed, the level of granularity of the sentence aligner is limited to words whether we consider them as sentences or not. We therefore another approach, using alignments between Inuktitut substrings and English tokens.

Martin et al. (2003) presented a study in building and using an English-Inuktitut parallel corpus. They described a sentence alignment technique tuned for the specificity of the Inuktitut language (namely, its agglutinative nature), and described as well a method for acquiring correspondent pairs of English tokens and Inuktitut substrings. The motivation behind their work was to populate a glossary with reliable such pairs. We

[2] `http://rali.iro.umontreal.ca/Japa`

adapted this approach in order to achieve word alignment.

4.1 Association Score

As Martin et al. (2003) pointed out, the strong agglutinative nature of Inuktitut makes it necessary to consider subunits of Inuktitut tokens. This is reflected by the large proportion of token types and hapax words observed on the Inuktitut side of the training corpus, compared to the ratios observed on the English side (see Table 1).

The main idea presented in (Martin et al., 2003) is to compute an association score between any English word seen in the English part of the training corpus and all the Inuktitut substrings of the tokens that were seen in the corresponding Inuktitut part of the corpus. To do so, they used a pointwise mutual information score. In our case, we computed a log-likelihood ratio score (Dunning, 1993) for all pairs of English tokens and Inuktitut substrings of length ranging from 3 to 10 characters. A maximum of L associations were kept for each English word (the top ranked ones) and then normalized such that for each English word e, we have a distribution of likely Inuktitut substrings s: $\sum_s p_{llr}(s|e) = 1$.

To reduce the computation load, we used a suffix tree structure and computed the association scores only for the English words belonging to the test corpus we had to align. We also filtered out Inuktitut substrings we observed less than three times in the training corpus. Altogether, it takes about one hour for a good desktop computer to produce the association scores for one hundred English words.

4.2 Word Alignment Strategy

Our approach for aligning an Inuktitut sentence of K tokens I_1^K with an English sentence of N tokens E_1^N (where $K \leq N$)[3] can be framed into seeking the best word alignment $\hat{A}$:

$$\begin{aligned} \hat{A} &= \text{argmax}_A\, P(A|I_1^K, E_1^N) \\ &= \text{argmax}_A\, P(I_1^K|E_1^N, A) \times P(A) \\ &\simeq \text{argmax}_A\, P(I_1^K|E_1^N) \times P(A) \end{aligned}$$

[3] As a matter of fact, the number of Inuktitut words in the test corpus is always less than or equal to the number of English tokens for any sentence pair.

We further make some assumptions on the nature of the alignment we seek, which greatly simplify the tractability of the model we propose. These assumptions rely on two observations we made on the manual alignment provided for the development set of the WPT'05 task (see Section 5.1): English-Inuktitut word alignment is almost monotonic and most of the time, one Inuktitut word is aligned to n adjacent English words.

Therefore, our maximization problem can be cast into finding $K-1$ *cutting points* $c_{k\in[1,K-1]}$ ($c_k \in [1, N-1]$) on the English side. A frontier c_k delimits adjacent English words $E_{c_{k-1}+1}^{c_k}$ that are translation of the single Inuktitut word I_k. Under an independence assumption of each alignment, and with the convention that $c_0 = 0$, $c_K = N$ and $c_{k-1} < c_k$, we can formulate our alignment problem as seeking the best word alignment $\hat{A}$ by maximizing:

$$\hat{A} = \underset{c_1^K}{\text{argmax}} \prod_{k=1}^{K} P(I_k|E_{c_{k-1}+1}^{c_k}) \times P(d_k)$$

where $d_k = c_k - c_{k-1}$ is the number of English words associated to I_k, $p(d_k)$ is the prior probability that d_k English words are aligned to a single Inuktitut word, which we computed directly from Table 2.

Note that our current implementation of this cutting point approach limits potential word alignments in two ways. First, we do not allow an Inuktitut word to be unaligned. That is, the condition $d_k > 0$ is always true for all alignments tried. This is not a big limitation, as an Inuktitut word is almost always quite long and the corresponding cept is bound to be aligned to at least one English word. Second, this approach cannot produce many-to-many word alignments, an English word is aligned to exactly one Inuktitut word (but an Inuktitut word may be aligned to many English words).

During the development cycle, we noticed slightly better results with this formulation:

$$\hat{A} = \underset{c_1^K}{\text{argmax}} \prod_{k=1}^{K} \alpha P(I_k|E_{c_{k-1}+1}^{c_k}) + (1-\alpha) P(d_k) \quad (1)$$

where α is a weighting coefficient. One reason for the slight gain in performance is that the prior $p(d_k)$ was only of little help.

4.3 Further Approximations

We tried the following two methods to approximate $p(I_k|E_{c_{k-1}+1}^{c_k})$. The second one led to better results.

$$p(I_k|E_{c_{k-1}+1}^{c_k}) \simeq \begin{cases} \max_{j=c_{k-1}+1}^{c_k} p(I_k|E_j) \\ \text{or} \\ \sum_{j=c_{k-1}+1}^{c_k} p(I_k|E_j) \end{cases}$$

We also considered several ways of computing the probability that an Inuktitut token I is the translation of an English one E; the best one we found being:

$$p(I|E) \simeq \sum_{s \in I} \lambda p_{llr}(s|E) + (1-\lambda) p_{ibm2}(s|E)$$

where the summation is taken over all subsequences of at least 3 characters of the Inuktitut word I, λ is a weighting coefficient, $p_{llr}(s|E)$ is the distribution described in Section 4.1 and $p_{ibm2}(s|E)$ is the probability obtained from an IBM model 2. Before training the IBM model 2, we segmented the Inuktitut words using a recursive procedure optimising a frequency-based criterion. This criterion seeks to maximize the product of the frequency of each segment with the constraint that each segment has a length of at least three characters.

We tried to directly embed a model trained on whole (unsegmented) Inuktitut tokens, but noticed a degradation in performance (line 2 of Table 4). Therefore, the segmented model was preferred throughout this study.

5 Experiments and Results

5.1 Corpora

All the corpora used in this study were provided as part of the WPT'05 material made available to participants (Martin et al., 2005). We worked on a collection of Inuktitut-English parallel texts from the Legislative Assembly of Nunavut, sentence-aligned. The alignment was provided by the Inuktitut Computing research team.[4]

Three corpora were provided: a training corpus (TRAIN), a development corpus (DEV) and, eventually, a test corpus (TEST) used to evaluate the participants. Some statistics on the training corpus are presented in Tables 1 and 2.

The development corpus provided for tuning our system was provided along with a word alignment manually built. It contains only 25 pairs of sentences. A test corpus of 75 pairs of sentences was also provided for which we had to provide an alignment. These three corpora do no overlap (no common pair of sentences). We tuned our systems using the DEV corpus and all our tests are made on the TEST corpus.

	Inuktitut	%	English	%
sentences	333 085		333 085	
tokens	2 153 034		3 992 298	
types	417 407	19.4	27 127	67.6
hapax	337 798	80.9	8 792	32.4

Table 1: Number of sentences and ratios of token types and hapax words (words seen only once) in the TRAIN corpus.

5.2 Metrics

We evaluated our alignments against a gold standard using *precision* (P), *recall* (R), *f-measure* (F) and *alignment error rate* (AER) (Och and Ney, 2000). Precision and AER measure the quality of the alignments produced, recall measures their coverage and f-measure synthesizes precision and recall. Those scores are defined by:

$$\begin{aligned} P &= \frac{|A \cap G|}{|A|} \\ R &= \frac{|A \cap G|}{|G|} \\ F &= \frac{2PR}{P+R} \\ AER &= 1 - \frac{|A \cap G_p| + |A \cap G_s|}{|A| + |G_s|} \end{aligned}$$

where A is the set of alignments returned by the system and G the set of alignments in the gold standard. G_s and G_p are respectively the subset of S(ure) and P(robable) alignments in G. According to Martin et al. (2005), whenever a single English word was aligned to a single Inuktitut word (whitespace-separated string), the alignment was qualified Sure. Otherwise, the Cartesian product of the aligned phrases (one of which being potentially a single word) was assigned a Probable tag. Roughly 14% of the alignments were Sure ones in the reference of the TEST corpus.

[4] See www.inuktitutcomputing.ca/NunavutHansard/

It is worth noting that since G_s is rather small, the AER is inversely correlated to the precision computed for Probable alignments.

5.3 Using a Sentence Aligner

Because in its default setting JAPA only considers `n-m` sentence alignment patterns with `n,m` $\in [0, 2]$, we provided it with a new pattern distribution which better fits the empirical one we observed on the DEV corpus (see Table 2).

1-1	0.406	4-1	0.092	2-2	0.000
2-1	0.172	5-1	0.04	3-2	0.011
3-1	0.123	6-1	0.04	4-2	0.015
7-2	0.011	7-1	0.027	5-2	0.011

Table 2: Distribution of the English-Inuktitut patterns given to JAPA.

We also set the number of English characters generated by an Inuktitut one to 1.05, a value which was observed in the TRAIN corpus. Surprisingly, although those two languages have very different word systems, they both use about the same amount of characters to express a message. Finally, we used a search space composed by a diagonal beam of a radius of 10 words.

`1-n` and `n-1` alignments identified by JAPA were output without further processing. Since the word alignment format of the shared task do not account for `n-m` alignments (`n,m` > 1) we generated the Cartesian product of the two sets of words for all these `n-m` alignments produced by JAPA.

The performances of the sentence aligner on the TEST corpus are reported in Table 3. We see that tuning is beneficial (line 3 *vs* line 1). At the same time, it is interesting to note that bad tuning is clearly disadvantageous (line 2). This bad performance was indeed the official one JAPA received at WPT (Langlais et al., 2005). The major difference between both tuning (line 2 and 3) is the pattern distribution we fed JAPA with. In the worst version, the distribution contains all the (24) `n-m` patterns observed on the DEV corpus. This has the undesirable effect that JAPA outputs frequently `n-m` patterns with `n` or `m` greater than unity, and therefore the Cartesian product we resort to in such cases drastically lowers the precision figure.

Configuration	P	R	F	AER
without tuning	53.04	37.12	43.68	45.13
WPT'05	26.17	74.49	38.73	71.27
after tuning	55.41	60.55	57.86	42.48

Table 3: Word alignment performances on the TEST corpus using our sentence aligner.

5.4 NUKTI

We quickly realised that, because of its combinatorial nature, the maximization of equation 1 was barely tractable. Therefore we adopted a greedy strategy to reduce the search space. We first computed a split of the English sentence into K adjacent regions c_1^K by virtually drawing a diagonal line we would observe if a character in one language was producing a constant number of characters in the other one. An initial word alignment was then found by simply tracking this diagonal at the word granularity level.

With this split in hand (line 1 in Table 4), we move each cutting point around its initial value starting from the leftmost cutting point and going rightward. Once a locally optimal cutting point is found (that is, maximizing the score of equation 1), we proceed to the next one, directly to its right.

We report in Table 4 the performance of three variants we tried. It is interesting to note that the starting point of the greedy search (line 1) does relatively well considering how simple the approach is. However, moving from this initial split clearly improves the performance (line 3).

variant	P	R	F	AER
start (diag)	54.20	56.59	55.37	45.54
greedy (word)	45.56	47.57	46.54	50.47
greedy (seg)	65.4	68.31	66.83	32.10

Table 4: Performance of several NUKTI alignment techniques measured on the DEV corpus. *start* is a simple diagonal, *greedy (word)* is the greedy search we describe in the text with an IBM model trained on an unsegmented corpus, *greedy (seg)* uses a IBM model trained on a segmented corpus.

We observed that putting much of the weight

λ on the IBM model 2 yielded the best results. This means that, to our utmost disappointment, the log-likelihood scores did little to improve the alignment quality.

One explanation for this could be that in the greedy variants reported in Table 4, we kept a maximum of $L = 25\,000$ Inuktitut associations for each English word. This may be too high: the worst-ranked ones are likely to be irrelevant and therefore part of the probability mass captured in p_{llr} may have been wasted on irrelevant substrings. Unfortunately, keeping only the $L = 200$ best associations (and then normalizing) only lead to a marginal improvement.

Also, during the tuning phase, we noticed that the prior $p(d_k)$ in equation 1 did not help (the factor $1-\alpha$ was close to 0). A character-based model might have been more appropriate to the case.

5.5 Combining JAPA and NUKTI

5.5.1 Avoiding Cartesian Products

One important weakness of our first approach lies in the Cartesian product we generate when JAPA produces a `n-m` ($n, m > 1$) alignment. Thus, we tried a third approach: we applied NUKTI on any `n-m` alignment JAPA produces as if this initial alignment were in fact two (small) sentences to align, `n`- and `m`-word long respectively. We can therefore avoid the Cartesian product and select word alignments more discerningly. As can be seen in Table 5, this combination improved over JAPA alone, while being worse than NUKTI alone (line 3 of Table 4).

variant	P	R	F	AER
JAPA	55.41	60.55	57.86	42.48
JAPA *no*-CP	57.36	59.89	58.60	40.73

Table 5: Performance of JAPA when avoiding Cartesian products with NUKTI as measured on the TEST corpus.

5.5.2 Using JAPA as a Seed

Another way to use JAPA with NUKTI is to consider the word alignment produced by JAPA as a the initial alignment (the *seed*). We took the best cutting points generated by JAPA and attempted once again to move them around their initial value to maximise equation 1, using the same procedure as the one described in Section 5.4.

Unfortunately, since, in its current implementation, NUKTI does not handle null alignments, we used as a seed the best result JAPA could give when constrained not to produce null alignments. The performance of this variant (line 1 of Table 6) is however not much lower than the JAPA variant (line 1 of Table 5).

The results are presented in Table 6. The improvement of the combination over JAPA alone is very significant: 10.93% absolute in AER. However, when compared to the best configuration of NUKTI (line 3 of Table 4), the improvement of 0.17% absolute in AER is much more modest. Nonetheless, this configuration is our best one.

variant	P	R	F	AER
JAPA *no-null*	55.0	60.1	57.48	42.85
JAPA+NUKTI	65.47	68.36	66.88	31.93

Table 6: Performance of NUKTI seeded with JAPA, measured on the TEST corpus.

5.6 Bias of our model

As already mentioned, both JAPA and NUKTI strongly rely on two assumptions: the monotonicity of the alignment, as well as the 1-to-N cardinality of the Inuktitut-English patterns. From the gold standard reference of the TEST material G (which was made available after the workshop), we computed several figures that help appreciate the bias of these hypotheses.

1.4% Inuktitut and 5.7% English words are left without any counterpart in G, something we can not handle presently with NUKTI. It is instructive to note that roughly 90% of the Inuktitut words are aligned to a sequence of adjacent English words; therefore we run over potential problems for 10% of the Inuktitut tokens we treat with NUKTI. Also, Schafer and Drábek (2005) computed that 4.7% of the Inuktitut positions associated to two adjacent English positions are in reverse order (a crossing that we cannot handle).

6 A look at WPT'05

Four teams (including ours) participated to the Inutktitut-English word alignment task of

WPT'05 (Schafer and Drábek, 2005; Lopez and Resnik, 2005; Caseli et al., 2005; Langlais et al., 2005).

Schafer and Drábek (2005) devised a weighted finite-state transducer (WSFT) which exploits the same properties of the alignment we capitalize on in this study, namely monotonicity and 1-to-N cardinality. This approach (line 1 of Table 7) compares impressively closely to the best variant we report in this study (line 2 of Table 6 also duplicated for convenience in the last line of Table 7).

They also report evaluations for alignments produced by several IBM model 4 trained with GIZA++: giza-I2E (resp. giza-E2I) is the alignment obtained when English (resp. Inuktitut) was the target language of the underlying model, giza-syll was produced after the Inuktitut material was first syllabized (with a segmenter they wrote within approximately 2 person-hours). One very interesting fact about their work is the voting procedure they tested which benefits from the various characteristics of their variants. They show that this allows to tune the final system to one of the metrics that is deemed important in a given situation. See (Martin et al., 2005) for the different tunings they submitted.

variant	P_P	R_P	F_P	AER
JHU-WFST	65.4	68.3	66.8	33.7
JHU-giza-I2E	49.7	18.6	27.0	45.2
JHU-giza-E2I	64.6	56.2	60.1	32.7
JHU-giza-syll	84.9	44.0	57.9	15.6
UMIACS	89.16	16.68	28.11	22.51
LIHLA	79.53	18.71	30.30	22.72
RALI	63.09	65.87	64.45	34.06
JAPA+NUKTI	65.47	68.36	66.88	31.93

Table 7: Precision, Recall, F-measure (measured on the P alignments) and AER obtained by different variants submitted at the WPT'05 shared-task. Individual variants of the JHU team are taken from (Schafer and Drábek, 2005). The last line of this table is only there for convenience and duplicates line 2 of Table 6.

Lopez and Resnik (2005) tested a refined version of the HMM alignment model embedded in GIZA++ (UMIACS). Last, Caseli et al. (2005) used a set of simple heuristics on top of two automatically computed bilingual lexicons ($I2E$ and $E2I$) to do the word alignment (LIHLA). Although different in nature, the last two systems show comparable performances that are characterized by a high precision but a fairly low recall.

Note that the figures of Table 7 illustrate the inverse correlation between AER and precision (P_p).

7 Conclusion

Within the framework of WPT'05, we proposed two methods for aligning an English-Inuktitut parallel corpus at the word level. The two months we had since the workshop that sparked our interest in Inuktitut have been quite productive. Many refinements were brought to the very core of NUKTI. Our best results were achieved when we seeded our NUKTI system with the results of JAPA, our sentence aligner.

Still, we believe that the NUKTI system could be improved. First, even if it has been shown that JAPA is performing well (Langlais et al., 1998), sentence-alignment technology evolved since then (Singh and Husain, 2005) and more recent aligners (*e.g.* (Moore, 2002)) might do better. Second, NUKTI has some intrinsic limitations, as for instance, the fact that it can align each English token with one and only one Inuktitut token. Indeed, had we had English sentences longer than their Inuktitut counterparts, we would not have been to handle them. Third, its greedy nature is very aggressive and only explore a small fraction of the full search space.

Another weakness of the approach using substrings of Inuktitut words is that the words are not necessarily segmented in a meaningful manner. We believe that NUKTI would greatly benefit from a Inuktitut morphological analyzer, which would help to identify meaningful substrings. At the time of this writing, such an analyzer is currently being worked on.[5]

Acknowledgements

We wish to thank the anonymous reviewers for the thorough comments they made on the first

[5]See `http://iit-iti.nrc-cnrc.gc.ca/projects-projets/uqausiit_e.html`

draft of this article. This work has been financially supported by grants from NSERC and FQRNT.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Helena M. Caseli, Maria G.V. Nunes, and Mikel L. Forcada. 2005. LIHLA: Shared-task system description. In *2nd ACL workshop on Building and Using Parallel Texts: Data Driven and Beyond (WPT)*, pages 111–114, Ann Arbor, Michigan, June 29-30.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1).

William A. Gale and Kenneth. W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. In *Computational Linguistics*, volume 19, pages 75–102.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL*, pages 273–280.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT*, pages 127–133.

Philippe Langlais, Michel Simard, and Jean Véronis. 1998. Methods and Practical Issues in Evaluating Alignment Techniques. In *36th annual meeting of the ACL*, Montreal, Canada.

Philippe Langlais, Fabrizio Gotti, and Guihong Cao. 2005. Nukti: English-inuktitut word-alignment system description. In *2nd WPT*, pages 75–78, Ann Arbor, Michigan, June 29-30.

Adam Lopez and Philip Resnik. 2005. Improved hmm alignment models for languages with scare resources. In *2nd WPT*, pages 83–86, Ann Arbor, Michigan, June 29-30.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and Using an English-Inuktitut Parallel Corpus. In *Building and using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 115–118, Edmonton, Canada.

Joel Martin, Rada Mihalcea, and Ted Perdersen. 2005. Word alignment for languages with scarce resources. In *2nd WPT*, pages 65–74, Ann Arbor, Michigan, June 29-30.

Robert C. Moore. 2002. Fast and accurate senetnce alignment of bilingual corpora. In *Proceedings of the 5th Associaton for Machine Translation in the Americas*, pages 135–144, Tiburon, California.

Franz Jospeh Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 1086–1090, Saarbrucken, Germany.

Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

Franz Joseph Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Compuutational Linguistics*.

Charles Schafer and Elliott Franco Drábek. 2005. Models for inuktitut-english word alignment. In *2nd WPT*, pages 79–82, Ann Arbor, Michigan, June 29-30.

Michel Simard, George Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.

Anil Kumar Singh and Samar Husain. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *2nd WPT*, pages 99–106, Ann Arbor, Michigan, June 29-30.

Ashish Venugopal, Stephan Vogel, and Alex Waibel. 2003. Effective phrase translation extraction from alignment models. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 319–326.

Bing Zhao and Stephan Vogel. 2005. A generalized alignment-free phrase extraction. In *2nd WPT*, pages 141–144, Ann Arbor, Michigan.