NUKTI: English-Inuktitut Word Alignment System Description

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

Abstract

Machine Translation (MT) as well as other bilingual applications strongly rely on word alignment. 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 effort on the English-Inuktitut word alignment shared task and report on two approaches we implemented and a combination of both.

1 Introduction

Word alignment is an important step in exploiting parallel corpora. When efficient techniques have been proposed (Brown et al., 1993; Och and Ney, 2003), they have been mostly evaluated on "safe" pairs of languages where the notion of word is rather clear.

We devoted two weeks to the intriguing task of aligning at the word level pairs of sentences of English and Inuktitut. We experimented with two different approaches. For the first one, we relied on an in-house sentence alignment program (JAPA) where English and Inuktitut tokens were considered as sentences. The second approach we propose takes advantage of associations computed between any English word and roughly any subsequence of Inuktitut characters seen in the training corpus. We also investigated the combination of both approaches.

2 JAPA: Word Alignment as a Sentence Alignment Task

To adjust our systems, the organizers made available to the participants a set of 25 pairs of sentences where words had been manually aligned. A fast inspection of this material reveals that in most of the cases, the alignment produced are monotonic and involve *cepts* of `n` adjacent English words aligned to a single Inuktitut word.

Many sentence alignment techniques strongly rely on the monotonic nature of the inherent alignment. Therefore, we conducted a first experiment using an in-house sentence alignment program called JAPA that we developed within the framework of the Arcade evaluation campaign (Langlais et al., 1998). The implementation details of this aligner can be found in (Langlais, 1997), but in a few words, JAPA aligns pairs of sentences by first grossly aligning their words (making use of either cognate-like tokens, or a specified bilingual dictionary). A second pass aligns the sentences in a way similar[1] to the algorithm described by Gale and Church (1993), but where the search space is constrained to be close to the one delimited by the word alignment. This technique happened to be among the most accurate of the ones tested during the Arcade exercise.

To adapt JAPA to our needs, we only did two things. First, we considered single sentences as documents, and tokens as sentences (we define a token as a sequence of characters delimited by

[1] In our case, the score we seek to globally maximize by dynamic programming is not only taking into account the length criteria described in (Gale and Church, 1993) but also a cognate-based one similar to (Simard et al., 1992).

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 75–78,
Ann Arbor, June 2005.

1-1	0.406	4-1	0.092	4-2	0.015
2-1	0.172	5-1	0.038	5-2	0.011
3-1	0.123	7-1	0.027	3-2	0.011

Table 1: The 9 most frequent English-Inuktitut patterns observed on the development set. A total of 24 different patterns have been observed.

white space). Second, since 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 we computed from the development corpus (see Table 1). It is interesting to note that although English and Inuktitut have very different word systems, the length ratio (in characters) of the two sides of the TRAIN corpus is 1.05.

Each pair of documents (sentences) were then aligned separately with JAPA. `1-n` and `n-1` alignments identified by JAPA where output without further processing. Since the word alignment format of the shared task do not account directly 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 performance of this approach is reported in Table 2. Clearly, the precision is poor. This is partly explained by the cartesian product we resorted to when `n-m` alignments were produced by JAPA. We provide in section 4 a way of improving upon this scenario.

Prec.	Rec.	F-meas.	AER
22.34	78.17	34.75	74.59

Table 2: Performance of the JAPA alignment technique on the DEV corpus.

3 NUKTI: Word and Substring Alignment

Martin et al. (2003) documented a study in building and using an English-Inuktitut bitext. They described a sentence alignment technique tuned for the specificity of the Inuktitut language, and described as well a technique 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 extended this line of work in order to achieve word alignment.

3.1 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 3).

	Inutktitut	%	English	%
tokens	2 153 034		3 992 298	
types	417 407	19.4	27 127	0.68
hapax	337 798	80.9	8 792	32.4

Table 3: Ratios of token types and happax words in the TRAIN corpus.

The main idea presented in (Martin et al., 2003) is to compute an association score between any English word seen in the training corpus and all the Inuktitut substrings of those tokens that were seen in the same region. In our case, we computed a 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 25 000 associations were kept for each English word (the top ranked ones).

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.

We normalize the association scores such that for each English word e, we have a distribution of likely Inuktitut substrings s: $\sum_s p_{llr}(s|e) = 1$.

3.2 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 $K \leq N$)[2] consists of finding

[2] 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.

$K-1$ *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. 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 $A = A(I_1^K|E_1^N)$ by maximizing:

$$A = \underset{c_1^K}{\operatorname{argmax}} \prod_{k=1}^{K} p(I_k|E_{c_{k-1}+1}^{c_k})^{\alpha_1} \times p(d_k)^{\alpha_2} \quad (1)$$

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 1; and α_1 and α_2 are two weighting coefficients.

We tried the following two approximations to compute $p(I_k|E_{c_{k-1}+1}^{c_k})$. 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 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 carried over all substrings s of I of 3 characters or more. $p_{llr}(s|E)$ is the normalized log-likelihood ratio score described above and $p_{ibm2}(s|E)$ is the probability obtained from an IBM model 2 we trained after the Inuktitut side of the training corpus was segmented using a recursive procedure optimizing a frequency-based criterion. λ is a weighting coefficient.

We tried to directly embed a model trained on whole (unsegmented) Inuktitut tokens, but noticed a degradation in performance (line 2 of Table 4).

3.3 A Greedy Search Strategy

Due to 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.

Having this split in hand (line 1 of 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 has been found (that is, maximizing the score of equation 1), we proceed to the next one directly to its right.

3.4 Results

We report in Table 4 the performance of different variants we tried as measured on the development set. We used these performances to select the best configuration we eventually submitted.

variant	Prec.	Rec.	F-m.	AER
start (diag)	51.7	53.66	52.66	49.54
greedy (word)	61.6	63.94	62.75	35.93
greedy (best)	63.5	65.92	64.69	34.21

Table 4: Performance of several NUKTI alignment techniques measured on the DEV corpus.

It is interesting to note that the starting point of the greedy search (line 1) does better than our first approach. However, moving from this initial split clearly improves the performance (line 3). Among the greedy variants we tested, we noticed that putting much of the weight λ on the IBM model 2 yielded the best results. We also noticed that $p(d_k)$ in equation 1 did not help (α_2 was close to zero). A character-based model might have been more appropriate to the case.

4 Combination of JAPA and NUKTI

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 apply 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.

5 Results

We submitted 3 variants to the organizers. The performances for each method are gathered in Table 5. The order of merit of each approach was consistent with the performance we measured on the DEV corpus, the best method being the NUKTI one. Curiously, we did not try to propose any Sure alignment but did receive a credit for it for two of the variants we submitted.

variant	T.	Prec.	Rec.	F-m.	AER
JAPA	P	26.17	74.49	38.73	71.27
JAPA +	S	9.62	67.58	16.84	
NUKTI	P	51.34	53.60	52.44	46.64
NUKTI	S	12.24	86.01	21.43	
	p	63.09	65.87	64.45	30.6

Table 5: Performance of the 3 alignments we submitted for the TEST corpus. $T.$ stands for the type of alignment (Sure or Possible).

6 Discussion

We proposed two methods for aligning an English-Inuktitut bitext at the word level and a combination of both. The best of these methods involves computing an association score between English tokens and Inuktitut substrings. It relies on a greedy algorithm we specifically devised for the task and which seeks a local optimum of a cumulative function of log-likelihood ratio scores. This method obtained a precision and a recall above 63% and 65% respectively.

We believe this method could easily be improved. First, it has some intrinsic limitations, as for instance, the fact that NUKTI only recognizes `1-n` cepts and do not handle at all unaligned words. Indeed, our method is not even suited to aligning English sentences with fewer words than their respective Inuktitut counterpart. Second, the greedy search we devised is fairly aggressive and only explores a tiny bit of the full search. Last, the computation of the association scores is fairly time-consuming.

Our idea of redefining word alignment as a sentence alignment task did not work well; but at the same time, we adapted poorly JAPA to this task. In particular, JAPA does not benefit here from all the potential of the underlying cognate system because of the scarcity of these cognates in very small sequences (words).

If we had to work on this task again, we would consider the use of a morphological analyzer. Unfortunately, it is only after the submission deadline that we learned of the existence of such a tool for Inuktitut[3].

Acknowledgement

We are grateful to Alexandre Patry who turned the JAPA aligner into a nicely written and efficient C++ program.

References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

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

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

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

P. Langlais. 1997. A System to Align Complex Bilingual Corpora. QPSR 4, TMH, Stockholm, Sweden.

J. Martin, H. Johnson, B. Farley, and A. 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.

F.J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

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

[3]See `http://www.inuktitutcomputing.ca/Uqailaut/`