Models for Inuktitut-English Word Alignment

Charles Schafer and Elliott Franco Drábek
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA
{*cschafer,edrabek*}*@cs.jhu.edu*

Abstract

This paper presents a set of techniques for bitext word alignment, optimized for a language pair with the characteristics of Inuktitut-English. The resulting systems exploit cross-lingual affinities at the sublexical level of syllables and substrings, as well as regular patterns of transliteration and the tendency towards monotonicity of alignment. Our most successful systems were based on classifier combination, and we found different combination methods performed best under the target evaluation metrics of F-measure and alignment error rate.

1 Introduction

Conventional word-alignment methods have been successful at treating many language pairs, but may be limited in their ability to generalize beyond the Western European language pairs for which they were originally developed, to pairs which exhibit more complex divergences in word order, morphology and lexical granularity. Our approach to Inuktitut-English alignment was to carefully consider the data in identifying difficulties particular to Inuktitut-English as well as possible simplifying assumptions. We used these observations to construct a novel weighted finite-state transducer alignment model as well as a specialized transliteration model. We combined these customized systems with 3 systems based on IBM Model 4 alignments under several methods of classifier combination. These combination strategies allowed us to produce multiple submissions targeted at the distinct evaluation measures via a precision/recall tradeoff.

2 Special Characteristics of the Inuktitut-English Alignment Problem

Guided by the discussion of Inuktitut in Mallon (1999), we examined the Nunavut Hansards training and hand-labeled trial data sets in order to identify special challenges and exploitable characteristics of the Inuktitut-English word alignment problem. We were able to identify three: (1) Importance of sublexical Inuktitut units; (2) 1-to-N Inuktitut-to-English alignment cardinality; (3) Monotonicity of alignments.

2.1 Types and Tokens

Inuktitut has an extremely productive agglutinative morphology, and an orthographic word may combine very many individual morphemes. As a result, in Inuktitut-English bitext we observe Inuktitut sentences with many fewer word tokens than the corresponding English sentences; the ratio of English to Inuktitut tokens in the training corpus is 1.85.[1] This suggests the importance of looking below the Inuktitut word level when computing lexical translation probabilities (or alignment affinities). To reinforce the point, consider that the ratio of training corpus types to tokens is 0.007 for English, and 0.194 for Inuktitut. In developing a customized word alignment solution for Inuktitut-English, a major goal was to handle the huge number of Inuktitut word types seen only once in the training corpus (337798 compared to 8792 for English), without demanding the development of a morphological analyzer.

2.2 Alignment

Considering English words in English sentence order, 4.7% of their alignments to Inuktitut were found to be *retrograde*; that is, involving a decrease in Inuktitut word position with respect to the previous English word's aligned Inuktitut position. Since this method of counting retrograde alignments would assign a low count to mass movements of large contiguous chunks, we also measured the number of inverted alignments over all pairs of English word positions. That is, the sum

$$\Sigma_e \Sigma_{a=1}^{a=|e|-1} \Sigma_{b=a+1}^{b=|e|} \Sigma_{i_1 \in I(e,a)} \Sigma_{i_2 \in I(e,b)} (1 \text{ if } i_1 > i_2)$$

was computed over all Inuktitut alignment sets $I(e, x)$, for e the English sentence and x the English word position. Dividing this sum by the obvious denominator (replacing $(1 \text{ if } i_1 > i_2)$ with (1) in the sum) yielded a value of 1.6% inverted alignments.

Table 1 shows a histogram of alignment cardinalities for both English and Inuktitut. Ninety-four percent of English word tokens, and ninety-nine percent of those having a non-null alignment, align to exactly one Inuktitut word. In development of a specialized word aligner for this language pair (Section 3), we made use of the observed reliability of these two properties, monotonicity and 1-to-N cardinality.

3 Alignment by Weighted Finite-State Transducer Composition

We designed a specialized alignment system to handle the above-mentioned special characteristics of Inuktitut-

[1] Though this ratio increases to 2.21 when considering only longer sentences (20 or more English words), ignoring common short, formulaic sentence pairs such as (Hudson Bay) (sanikiluaq) .

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 79–82,
Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

	% Words Having Specified Alignment Cardinality							
	NULL	**1**	**2**	**3**	**4**	**5**	**6**	**7**
English	5	94	<1	<1	0	0	0	0
Inuktitut	3	43	20	14	10	5	3	2

Table 1: Alignment cardinalities for English-Inuktitut word alignment, computed over the trial data.

English alignment. Our weighted finite-state transducer (WFST) alignment model, illustrated in Figure 1, structurally enforces monotonicity and 1-to-N cardinality, and exploits sublexical information by incorporating association scores between English words and Inuktitut word substrings, based on co-occurrence in aligned sentences. For each English word, an association score was computed not only with each Inuktitut word, but also with each Inuktitut character string of length ranging from 2 to 10 characters. This is similar to the technique described in Martin et al. (2003) as part of their construction of a bilingual glossary from English-Inuktitut bitext. However, our goal is different and we keep *all* the English-Inuktitut associations, rather than selecting only the "best" ones using a greedy method, as do they. Additionally, before extracting all substrings from each Inuktitut word, we added a special character to the word's beginning and end (e.g., *makkuttut* → *_makkuttut_*), in order to exploit any preferences for word-initial or -final placement.

The heuristic association score chosen was $p(word_e|word_i) \times p(word_i|word_e)$, computed over all the aligned sentence pairs. We have in the past observed this to be a useful indicator of word association, and it has the nice property of being in the range (0,1].

The WFST aligner is a composition of 4 transducers.[2] The structure of the entire WFST composition enforces monotonicity, Inuktitut-to-English 1-N cardinality, and Inuktitut word fertilities ranging between 1 and 7. This model was implemented using the ATT finite-state toolkit (Mohri et al., 1997). In Figure 1, **[1]** is a linear transducer mapping each English position in a particular English test sentence to the word at that position. It is constructed so as to force each English word to participate in exactly 1 alignment. **[2]** is a single-state transducer mapping English word to Inuktitut substrings (or full words) with weights derived from the association scores.[3] **[3]** is a transducer mapping Inuktitut substrings (and full words) to their position in the Inuktitut test sentence. Its construction allows a single Inuktitut position to correspond to multiple English positions, while enforcing monotonicity. **[4]** is a transducer regulating the allowed "fertility" values of Inuktitut words; each Inuktitut word is permitted a fertility of between 1 and 7. The fertility values are assigned the probabilities corresponding to observed relative frequencies in the *trial* data, and are not conditioned on the identity of the Inuktitut word.

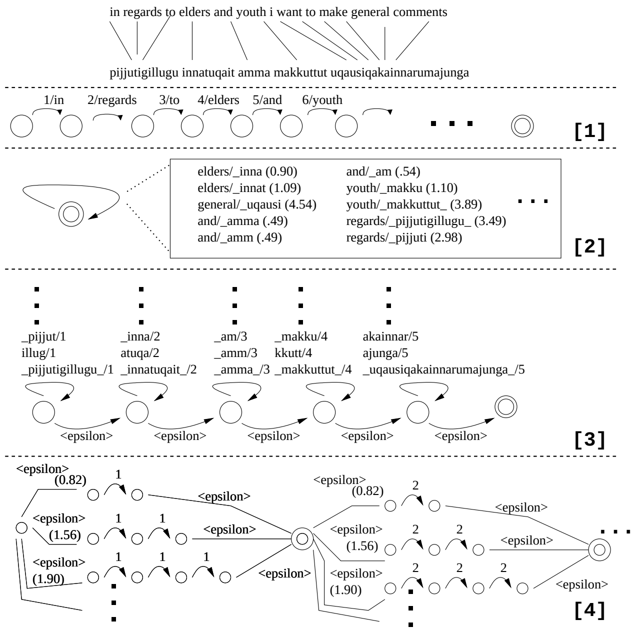

Figure 1: WFST alignment system in composition order, instantiated for an example sentence from the development (trial) data. To save space, only a representative portion of each machine is drawn. Transition weights are costs in the tropical **(min,+)** semiring, derived from negative logs of probabilities and association scores. Nonzero costs are indicated in parentheses.

4 English-Inuktitut Transliteration

Although in this corpus English and Inuktitut are both written in Roman characters, English names are significantly transformed when rendered in Inuktitut text. Consider the following English/Inuktitut pairs from the training corpus: **Chartrand/saaturaan**, **Chretien/kurittian** and the set of training corpus-attested Inuktitut renderings of **Williams**, **Campbell**, and **McLean** shown in Table 2(A) (which does not include variations containing the common **-mut** lexeme, meaning "to [a person]" (Mallon, 1999)).

Clearly, not only does the English-to-Inuktitut transformation radically change the name string, it does so in a nondeterministic way which appears to be influenced not only by the phonological preferences of Inuktitut but also by differing pronunciations of the name in question and possibly by differing conventions of translators (note, for example, **maklain** versus **mikliin** for **McLean**).

We trained a probabilistic finite-state transducer (FST) to identify English-Inuktitut transliterated pairs in aligned sentences. Training string pairs were acquired from the training bitext in the following manner. Whenever single instances of corresponding honorifics were found in a sentence pair – these included the correspondences (Ms , mis); (Mrs , missa/missis); (Mr ,

[2] Bracketed numbers in the following discussion refer to the component transducers as illustrated in Figure 1.

[3] Transducers **[2]** and **[4]** are shared across all sentence decodings.

(A)		(B)			
Williams	**McLean**	**k**		**sh**	
ailiams	makalain	k	-4.2	s	-7.2
uialims	makkalain	q	-6.2		
uilialums	maklaain			**w**	
uiliam	maklain	**b**		ui	-5.8
uiliammas	maklainn	p	-4.3	v	-6.1
uiliams	maklait	v	-5.0		
uilians	makli			**o**	
uliams	maklii	**z**		a	-4.2
viliams	makliik	j	-5.2	aa	-4.6
	makliin	s	-5.8	uu	-4.9
Campbell	maklin			u	-5.1
kaampu	malain	**ch**			
kaampul	matliin	s	-5.6	**u**	
kaamvul	miklain	k	-6.8	uu	-5.5
kamvul	mikliin			u	-5.6
	miklin			a	-6.2

Table 2: (A) Training-corpus-attested renderings of **Williams**, **Campbell**, and **McLean**. (B) Top learned Inuktitut substitutions and their log probabilities for several English *(shown underlined)* orthographic characters (and character sequences). Where top substitutions for English characters are shown, none equal or better were omitted.

mista/mistu) – the immediately following capitalized English words (up to 2) were extracted and the same number of Inuktitut words were extracted to be used as training pairs. Thus, given the appearance in aligned sentences of “Mr. Quirke” and “mista kuak”, the training pair (Quirke,kuak) would be extracted. Common distractions such as “Mr Speaker” were filtered out. In order to focus on the native English name problem (Inuktitut name rendering into English is much less noisy) the English extractions were required to have appeared in a large, news-corpus-derived English wordlist. This procedure resulted in a conservative, high-quality list of 434 unique name pairs. The probabilistic FST model we selected was that of a memoryless (single-state) transducer representing a joint distribution over character substitutions, English insertions, and Inuktitut insertions. This model is identical to that presented in Ristad and Yianilos (1997). Prior to training, common English digraphs (e.g., “th” and “sh”) were mapped to unique single characters, as were doubled consonants. Inuktitut “ng” and common two-vowel sequences were also mapped to unique single characters to elicit higher-quality results from the memoryless transduction model employed. Some results of the transducer training are displayed in Table 2(B). Probabilistic FST weight training was accomplished using the Dyna modeling language and DynaMITE parameter optimization toolkit (Eisner et al, 2004). The transliteration modeling described here differs from such previous transliteration work as Stalls and Knight (1998) in that there is no explicit modeling of pronunciation, only a direct transduction between written forms.

In applying transliteration on trial/test data, the following criteria were used to select English words for transliteration: *(1) Word is capitalized (2) Word is not in the exclusion list.*[4] For the top-ranked transliteration of the English word present in the Inuktitut sentence, all occurrences of that word in that sentence are marked as aligned to the English word.

We have yet to evaluate English-Inuktitut transliteration in isolation on a large test set. However, accuracy on the workshop trial data was 4/4 hypotheses correct, and on test data 2/6 correct. Of the 4 incorrect test hypotheses, 2 were mistakes in identifying the correct transliteration, and 2 mistakes resulted from attempting to transliterate an English word such as “Councillors” which should not be transliterated. Even with a relatively low accuracy, the transliteration model, which is used only as an individual voter in combination systems, is unlikely to vote for the incorrect choice of another system. Its purpose under system combination is to push a good alignment link hypothesis up to the required vote threshold.[5]

5 IBM Model 4 Alignments

As a baseline and contributor to our combination systems, we ran GIZA++ (Och and Ney, 2000), to produce alignments based on IBM Model 4. The IBM alignment models are asymmetric, requiring that one language be idenitifed as the “e” language, whose words are allowed many links each, and the other as the “f” language, whose words are allowed at most one link each. Although the observed alignment cardinalities naturally suggest identifying Inuktitut as the “e” language and English as the “f” language, we ran both directions for completeness.

As a crude first attempt to capture sublexical correspondences in the absence of a method for morpheme segmentation, we developed a rough syllable segmenter (spending approximately 2 person-hours), ran GIZA++ to produce alignments treating the syllables as words, and chose, for each English word, the Inuktitut word or words the largest number of whose syllables were linked to it.

In the nomenclature of our results tables, **giza++ syllabized** refers to the latter system, **giza++ E(1)-I(N)** represents GIZA++ run with English as the “e” language, and **giza++ E(N)-I(1)** sets English as the “f” language.

6 System Performance and Combination Methods

We observed the 4 main systems (3 GIZA++ variants and WFST) to have significantly different performance profiles in terms of precision and recall. Consistently, WFST

[4] Exclusion list was compiled as follows: (a) capitalized words in 2000 randomly selected English training sentences were examined, Words such as *Clerk*, *Federation*, and *Fisheries*, which are frequently capitalized but should not be transliterated, were put into the exclusion list; in addition, any word with frequency $>$ 50 in the training corpus was excluded, on the rationale that common-enough words would have well-estimated translation probabilities already. 50 may seem like a high threshold until one considers the high variability of the transliteration process as demonstrated in Table 2(A).

[5] Refer to Section 6 for detailed descriptions of voting.

SYSTEM	P	R	F	AER	$\lvert H\rvert/\lvert T\rvert$
Individual system performance **Trial Data**					
giza++ E(1)-I(N)	63.4	26.6	37.5	32.9	0.42
giza++ E(N)-I(1)	68.2	59.4	63.5	28.6	0.87
giza++ syllabized	83.6	44.5	58.1	**18.3**	0.53
WFST	70.3	72.7	**71.5**	27.8	1.03
Combination system performance **Trial Data**					
F/AER Emphasis	85.4	63.5	72.9	12.3	0.74
AER Emphasis (1)	92.6	44.2	59.9	**8.8**	0.48
AER Emphasis (2)	95.1	38.0	54.3	9.5	0.40
F Emphasis	74.8	77.6	**76.2**	21.9	1.04
Recall Emphasis	66.9	82.1	73.8	28.9	1.23
Individual system performance **Test Data**					
giza++ E(1)-I(N)	49.7	18.6	27.0	45.2	0.37
giza++ E(N)-I(1)	64.6	56.2	60.1	32.7	0.87
giza++ syllabized	84.9	44.0	57.9	**15.6**	0.52
WFST	65.4	68.3	**66.8**	33.7	1.04
(submitted) *Combination system performance* **Test Data**					
F/AER Emphasis	84.4	58.6	69.2	14.3	0.69
AER Emphasis (1)	90.7	39.4	54.9	11.5	0.43
AER Emphasis (2)	96.7	32.3	48.4	**9.5**	0.33
F Emphasis	70.7	73.8	**72.2**	26.7	1.04
Recall Emphasis	62.6	81.7	70.1	34.2	1.31

Table 3: System performance evaluated on trial and test data. The precision, recall and F-measure cited are the unlabeled version ("probable," in the nomenclature of this shared task). The gold standard truth for trial data contained 710 alignments. The test gold standard included 1972 alignments. The column $|H|/|T|$ lists ratio of hypothesis set size to truth set size for each system.

won out on F-measure while **giza++ syllabized** attained better alignment error rate (AER). Refer to Table 3 for details of performance on trial and test data.

We investigated a number of system combination methods, three of which were finally selected for use in submitted systems. There were two basic methods of combination: *per-link voting* and *per-English-word* voting.[6] In per-link voting, an alignment link is included if it is proposed by at least a certain number of the participating individual systems. In per-English-word voting, the best outgoing link is chosen for each English word (the link which is supported by the greatest number of individual systems). Any ties are broken using the WFST system choice. A high-recall variant of per-English-word voting was included in which ties at vote-count 1 (indicating a low-confidence decision) are not broken, but rather all systems' choices are submitted as hypotheses.

The transliteration model described in Section 4 was included as a voter in each combination system, though it made few hypotheses (6 on the test data). Composition of the submitted systems was as follows: **F/AER Emphasis** - per-link voting with decision criterion $>=$ 2 votes, over all 5 described systems (WFST, 3 GIZA++ variants, transliteration). **AER Emphasis (I)** per-link voting, $>=$ 2 votes, over all systems except giza++ E(N)-I(1). **AER Emphasis (II)** per-link voting, $>=$ 3 votes, over all systems. **F Emphasis** per-English-word voting, over all systems, using WFST as tiebreaker. **Recall Emphasis** per-English-word voting, over all systems, high-recall variant.

We elected to submit these systems because each tailors to a distinct evaluation criterion (as suggested by the naming convention). Experiments on trial data convinced us that minimizing AER and maximizing F-measure in a single system would be difficult. Minimizing AER required such high-precision results that the tradeoff in recall greatly lowered F-measure. It is interesting to note that system combination does provide a convenient means for adjusting alignment precision and recall to suit the requirements of the problem or evaluation standard at hand.

7 Conclusions

We have presented several individual and combined systems for word alignment of Inuktitut-English bitext. The most successful individual systems were those targeted to the specific characteristics of the language pair. The combined systems generally outperformed the individual systems, and different combination methods were able to optimize for performance under different evaluation metrics. In particular, per-English-word voting performed well on F-measure, while per-link voting performed well on AER.

Acknowledgements: *Many thanks to Eric Goldlust, David Smith, and Noah Smith for help in using the* Dyna *language.*

References

J. Eisner, E. Goldlust, and N. A. Smith. 2004. Dyna: A declarative language for implementing dynamic programs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Companion Volume, pages 218-221.

M. Mallon. 1999. Inuktitut linguistics for technocrats. Technical report, Ittukuluuk Language Programs, Iqaluit, Nunavut, Canada.

J. Martin, H. Johnson, B. Farley, and A. Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, HLT-NAACL 2003.

M. Mohri, F. Pereira, and M. Riley. 1997.
ATT General-purpose finite-state machine software tools.
http://www.research.att.com/sw/tools/fsm/.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

E. S. Ristad and P. N. Yianilos. 1997. Learning string edit distance. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 287–295.

B. Stalls and K. Knight. 1998. Translating names and technical terms in arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages.*

[6] Combination methods we elected not to submit included voting with trained weights and various stacked classifiers. The reasoning was that with such a small development data set – 25 sentences – it was unsafe to put faith in any but the simplest of classifier combination schemes.