Aligning and Using an English-Inuktitut Parallel Corpus

Joel Martin, Howard Johnson, Benoit Farley and Anna Maclachlan
Institute for Information Technology
National Research Council Canada
firstname.lastname@nrc-cnrc.gc.ca

Abstract

A parallel corpus of texts in English and in Inuktitut, an Inuit language, is presented. These texts are from the Nunavut Hansards. The parallel texts are processed in two phases, the sentence alignment phase and the word correspondence phase. Our sentence alignment technique achieves a precision of 91.4% and a recall of 92.3%. Our word correspondence technique is aimed at providing the broadest coverage collection of reliable pairs of Inuktitut and English morphemes for dictionary expansion. For an agglutinative language like Inuktitut, this entails considering substrings, not simply whole words. We employ a Pointwise Mutual Information method (PMI) and attain a coverage of 72.3% of English words and a precision of 87%.

1 Introduction

We present an aligned parallel corpus of Inuktitut and English from the Nunavut Hansards. The alignment at the sentence level and the word correspondence follow techniques described in the literature with augmentations suggested by the specific properties of this language pair. The lack of lexical resources for Inuktitut, the unrelatedness of the two languages, the fact that the languages use a different script and the richness of the morphology in Inuktitut have guided our choice of technique. Sentences have been aligned using the length-based dynamic programming approach of Gale and Church (1993) enhanced with a small number of lexical and non-alphabetic anchors. Word correspondences have been identified with the goal of finding an extensive high quality candidate glossary for English and Inuktitut words. Crucially, the algorithm considers not only full word correspondences, as most approaches do, but also multiple substring correspondences resulting in far greater coverage.

2 An English-Inuktitut Corpus

2.1 The Parallel Texts

The corpus of parallel texts we present consists of 3,432,212 words of English and 1,586,423 words of Inuktitut from the Nunavut Hansards. These Hansards are available to the public in electronic form in both English and Inuktitut (www.assembly.nu.ca). The Legislative Assembly of the newly created territory of Nunavut began sitting on April 1, 1999. Our corpus represents 155 days of transcribed proceedings of the Nunavut Legislative Assembly from that first session through to November 1, 2002, which was part way through the sixth session of the assembly.

We gather and process these 155 documents in various ways described in the rest of this paper and make available a sentence-aligned version of the parallel texts (www.InuktitutComputing.ca/NunavutHansards). Like the French-English Canadian Hansards of parliamentary proceedings, this corpus represents a valuable resource for Machine Translation research and corpus research as well as for the development of language processing tools for Inuktitut. The work reported here takes some first steps toward these ends, and it is hoped that others will find ways to expand on this work. One reason that the Canadian Hansards, a large parallel corpus of English-French, are particularly useful for research is that they are comparatively noise free as parallel text collections go (Simard and Plamondon, 1996). This should be true of the Nunavut Hansard collection as well. The Canadian Hansard is transcribed in both languages so what was said in English is transcribed in English and then translated into French and vice versa. For the Nunavut Hansard, in contrast, a complete English version of the proceedings is prepared and then this is translated into Inuktitut, even when the original proceedings were spoken in Inuktitut.

2.2 The Inuktitut Language

Inuktitut is the language of the Inuit living in North Eastern Canada, that is, Nunavut (Keewatin and Baffin Island), Nunavik and Labrador. It includes six closely related spoken dialects: Kivalliq, Aivilik, North Baffin, South Baffin, Arctic Quebec (Nunavik), and Labrador.

Inuktitut is a highly agglutinative language. Noun and verb roots occur with two main types of suffixes and there are many instantiations of these suffixes. The semantic suffixes modify the meaning of the root (over 250 of these in North Baffin dialect) and the grammatical suffixes indicate features like agreement and mood (approximately 700 verbal endings and over 300 nominal endings in North Baffin dialect).

A single word in Inuktitut is often translated with multiple English words, sometimes corresponding to a full English clause. For example, the Inuktitut word ᖃᐃᓵᓕᓂᐊᖅᑰᖖᒋᒃᑲᓗᐊᖅᐳᖅ (which is transliterated as *qaisaaliniaqquunngikkaluaqpuq*) corresponds to these eight English words: 'Actually he will probably not come early today'. The verbal root is *qai* 'come', the semantic suffixes are *-saali-*, *-niaq-*, *-qquu-*, *-nngit-* and *-galuaq-* meaning 'soon', 'a little later today or tomorrow', 'probability', 'negation', and 'actuality' respectively, and finally the grammatical suffix *-puq* expresses the 3rd person singular of the indicative mood. This frequently occurring one-to-many correspondence represents a challenge for word correspondence. The opposite challenging situation, namely instances of many-to-one correspondences, also arises for this language pair but less frequently. The latter is therefore not addressed in this paper.

Yet another challenge is the morphophonological complexity of Inuktitut as reflected in the orthography, which has two components. First, the sheer number of possible suffixes mentioned above is problematic. Second, the shape of these suffixes is variable. That is, there are significant orthographic changes to the individual morphemes when they occur in the context of a word. This type of variability can be seen in the above example at the interface of *-nngit-* and *-galuaq-*, which together become *-nngikkaluaq-*.

Finally, it is important to note that Inuktitut has a syllabic script for which there is a standard Romanization. To give an idea of how the scripts compare, our corpus of parallel texts consists of 20,124,587 characters of English and 13,457,581 characters in Inuktitut syllabics as compared to 21,305,295 characters of Inuktitut in Roman script.

3 Sentence Alignment

3.1 Sentence Alignment Approach

The algorithm used to align English-Inuktitut sentences is an extension of that presented in Gale and Church (1993). It does not identify crossing alignments where the sentence order within paragraphs in the parallel texts differs. Sentence alignments typically involve one English sentence matching one Inuktitut sentence (a 1-to-1 bead), but may also involve 2-to-1, 1-to-2, 0-to-1, 1-to-0 and 2-to-2 sentence matching patterns, or beads. Using such a length-based approach where the length of sentences is measured in characters is appropriate for our language pair since the basic assumption generally holds. Namely, longer English sentences typically correspond to longer Inuktitut sentences as measured in characters.

One problem with the approach, as pointed out by Macklovitch and Hannan (1998), is that from the point where a paragraph is misaligned, it is difficult to ensure proper alignment for the remainder of the paragraph. We observed this effect in our alignment. We also observed that the large number of small paragraphs with almost identical length caused problems for the algorithm.

Many alignment approaches have addressed such problems by making use of additional linguistic clues specific to the languages to be aligned. For our language pair, it was not feasible to use most of these. For example, some alignment techniques make good use of cognates (Simard and Plamondon, 1996). The assumption is that words in the two languages that share the first few letters are usually translations of each other. English and Inuktitut, however, are too distantly related to have many cognates. Even the translation of a proper name does not usually result in a cognate for our language pair, since the translation between scripts induces a phonetic translation rather than a character-preserving translation of the name, as these pairs illustrate *Peter*, *Piita*; *Canada*, *Kanata*; *McLean*, *Makalain*.

Following a suggestion in Gale and Church (1993), the alignment was aided by the use of additional anchors that were available for the language pair. These anchors consisted of non-alphabetic sequences (such as `9:00`, `42-1(1)` and `1999`) and 8 reliable word correspondences that occurred frequently in the corpus, including words beginning with these character sequences `speaker/uqaqti` and `motion/pigiqati`, for example.

3.2 Steps in Sentence Alignment

Preprocessing: Preprocessing the Inuktitut and the English raised separate issues. For English, the main issue was ensuring that illegal or unusual characters are mapped to other characters to simplify later processing. For Inuktitut the main issue was the array of encodings used for the syllabic script. Inuktitut syllabics can be represented using a 7-bit encoding called *ProSyl*, which is in many cases extended to an 8-bit encoding *Tunngavik*. Each syllabic character can be encoded in multiple ways that need to be mapped into a uniform scheme, such as

Unicode. Each separate file was converted to HTML using a commercial product *LogicTran r2net*. Then, the Perl package HTML::TreeBuilder was used to purge the text of anomalies and set up the correct mappings. The output of this initial preprocessing step was a collection of HTML files in pure Unicode UTF8.

Boundary Identification: The next step was to identify the paragraph and sentence boundaries for the Inuktitut and English texts. Sentences were split at periods, question marks, colons and semi-colons except where the following character was a lower case letter or a number. This resulted in a number of errors but was quite accurate in general. Paragraph boundaries were inserted where such logical breaks occurred as signaled in the HTML and generally correspond to natural breaks in the original document. Using HTML indicators contributed to the number of very short paragraphs, especially toward the beginning of each document. As mentioned in section 3.1, these short paragraphs were problematic for the alignment algorithm. The collection consists of 348,619 sentences in 112,346 paragraphs in English and 352,486 sentences in 118,733 paragraphs in Inuktitut. After this step, document, paragraph and sentence boundaries were available to use as hard and soft boundaries for the Gale and Church algorithm.

Syllabic Script Conversion: The word correspondence phase required a Roman script representation of the Inuktitut texts. The conversion from unicode syllabics to Roman characters was performed at this stage in the sentence alignment process using the standard ICI conversion method.

Anchors: The occurrences of the lexical anchors mentioned above were found and used with a dynamic programming search to find the path with the largest number of alignments. This algorithm was written in Perl and required about two hours to process the whole corpus. All alignments that occurred in the first two sentences of each paragraph were marked as hard boundaries for the Gale and Church (1993) program as provided in their paper.

3.3 Sentence Alignment Evaluation

Three representative days of Hansard (1999/04/01, 2001/02/21 and 2002/10/29) were selected and manually aligned at the sentence level as a gold standard. Precision and recall were then measured as suggested in Isabelle and Simard (1996).

Results: The number of sentence alignments in the gold standard was 3424. The number automatically aligned by our method was 3459. The number of those automatic alignments that were correct as measured against the gold standard was 3161. This represents a precision of 91.4% and a recall rate of 92.3%. For comparison, the Gale and Church (1993) program, which did not make use of additional anchors, had poorer results over our corpus. Their one-pass approach, which ignores paragraph boundaries, had a precision of 66.7% and a recall of 71.5%. Their two-pass approach, which aligns paragraphs in one pass and then aligns sentences in a second pass, had a precision of 85.6% and a recall of 87.0%.

4 Word Correspondence

Having built a sentence-aligned parallel corpus, we next attempted to use that corpus. Our goal was to extract as many reliable word associations as possible to aid in developing a morphological analyzer and in expanding Inuktitut dictionaries. The output of this glossary discovery phase is a list of suggested pairings that a human can consider for inclusion in a dictionary. Inuktitut dictionaries often disagree because of spelling and dialectical differences. As well, many contemporary words are not in the existing dictionaries. The parallel corpus presented here can be used to augment the dictionaries with current words, thereby providing an important tool for students, translators, and others.

In our approach, a glossary is populated with pairs of words that are consistent translations of each other. For many language pairs, considering whole word to whole word correspondences for inclusion in a glossary would yield good results. However, because Inuktitut is agglutinative, the method must discover pairs of an English word and the corresponding root of the Inuktitut word, or the corresponding Inuktitut suffix, or sometimes the whole Inuktitut word. In other words, it is essential to consider substrings of words for good coverage for a language pair like ours.

4.1 Substring Correspondence Method

Searching for substring correspondences is reduced to a counting exercise. For any pair of substrings, you need to know how many parallel regions contained the pair, how many regions in one language contained the first, how many regions in the other language contained the second, and how many regions there are in total. For example, the English word 'today' and the Inuktitut word 'ullumi' occur in 2092 parallel regions. The word 'today' appears in a total of 3065 English regions; and 'ullumi' appears in 2702 Inuktitut regions. All together, there are 332,154 aligned regions. It is fairly certain that these two words should be a glossary pair because each usually occurs as a translation of the other.

The PMI Measure: We measure the degree of association between any two substrings, one in the English and one in the Inuktitut, using Pointwise Mutual Information (PMI). PMI measures the amount of information that each substring conveys about the occurrence of the other. We recognize that PMI is badly behaved when the counts are near 1. To protect against that problem, we compute the 99.99999% confidence intervals around the

PMI (Lin, 1999), and use the lower bound as a measure of association. This lower bound rises as the PMI rises or as the amount of data increases. Many measures of association would likely work as well as the lower confidence bound on PMI. We used that bound as a metric in this study for three reasons. First, that metric led to better performance than Chi-squared on this data. Second, it addressed the problem of low frequency events. Third, it makes the correct judgment on Gale and Church's well-known chambre-communes problem (Gale and Church, 1991).

The decision to include pairs of substrings in the glossary proceeds as follows. Include the highest PMI scoring pairs if neither member of the pair has yet been included. If two pairs are tied, check whether the Inuktitut members of the pairs are in a substring relation. If they are, then add the pair with the longer substring to the glossary; if not, then add neither pair.

Many previous efforts have used a similar methodology but were only able to focus on word to word correspondences (Gale and Church, 1991). Here, the English words can correspond to any substring in any Inuktitut word in the aligned region. This means that statistics have to be maintained for many possible pairs. Under our approach, we maintain all these statistics for all English words, all Inuktitut words as well as substrings with length of between one and 10 Roman characters, and all co-occurrences that have frequency greater than three. This approach thereby addresses the challenge of Inuktitut roots and multiple semantic suffixes corresponding to individual English words. It also addresses the challenge of orthographic variation at morpheme boundaries to some degree since it will truncate morphemes appropriately in many cases.

4.2 Glossary Evaluation

This method suggested 4362 word-substring pairs for inclusion in a glossary. This represents a 72.3% coverage of English word occurrences in the corpus (omitting words of fewer than 3 characters). One hundred of these word-substring pairs were chosen at random and judged for accuracy using two existing dictionaries and a partial suffix list. An Inuktitut substring was said to match an English word *exactly* if the Inuktitut root plus all the suffixes carried the same meaning as the English word and conveyed the same grammatical features (e.g., grammatical number and case). The correspondence was said to be *good* if the Inuktitut root plus the left-most lexical suffixes conveyed the same meaning as the English word. In those cases, the Inuktitut word conveyed additional semantic or grammatical information.

About half of the *exact* matches were uninflected proper nouns. A typical example of the other exact matches is the pair *inuup* and *person's*. In this pair, *inu-* means person and *-up* is the singular genitive case. A typical example of a *good* match is the pair *pigiaqtitara* and *deal*. In this pair, *pigiaqti-* means deal and *-tara* conveys first person singular subject and third person singular object. For example, "I deal with him".

Of the 100 pairs, 43 were deemed exact matches and 44 were deemed good matches. The remaining 13 were incorrect. Taken together 87% of the pairs in the sample were useful to include in a glossary. This level of performance will improve as we introduce morphological analysis to both the Inuktitut and English words.

5 Conclusion

We have shown that aligning an English text with a highly agglutinative language text can have very useful outcomes. The alignment of the corpus to the sentence level was achieved accurately enough to build a usable parallel corpus. This is demonstrated by the fact that we could create a glossary tool on the basis of this corpus that suggested glossary pairings for 72.3% of English words in the text with a precision of 87%. We hope that our work will generate further interest in this newly available English-Inuktitut parallel corpus.

Acknowledgements We would like to thank Gavin Nesbitt of the Legislative Assembly of Nunavut for providing the Hansards, Peter Turney for useful suggestions, and Canadian Heritage for financial support of this project.

References

William A. Gale and Kenneth Ward Church. 1991. Identifying word correspondance in parallel text. In *Proceedings of the DARPA NLP Workshop.*

William A. Gale and Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–103.

Pierre Isabelle and Michel Simard. 1996. Propositions pour la représentation et l'évaluation des alignements de textes parallèles. [http://www.lpl.univ-aix.fr/projects/arcade/2nd/sent/metrics.html]. In *Rapport technique, CITI.*

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the ACL.*

Elliot Macklovitch and Marie-Louise Hannan. 1998. Line 'em up: Advances in alignment technology and their impact on translation support tools. *Machine Translation*, 13(1).

Michel Simard and Pierre Plamondon. 1996. Bilingual sentence alignment: Balancing robustness and accuracy. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA).*