Evaluating a Morphological Analyser of Inuktitut

Jeremy Nicholson†, Trevor Cohn‡ and Timothy Baldwin†
†Department of Computing and Information Systems, The University of Melbourne, Australia
‡Department of Computer Science, The University of Sheffield, UK
jeremymn@csse.unimelb.edu.au, tcohn@dcs.shef.ac.uk, tb@ldwin.net

Abstract

We evaluate the performance of an morphological analyser for Inuktitut across a medium-sized corpus, where it produces a useful analysis for two out of every three types. We then compare its segmentation to that of simpler approaches to morphology, and use these as a pre-processing step to a word alignment task. Our observations show that the richer approaches provide little as compared to simply finding the head, which is more in line with the particularities of the task.

1 Introduction

In this work, we evaluate a morphological analyser of Inuktitut, whose polysynthetic morphosyntax can cause particular problems for natural language processing; but our observations are also relevant to other languages with rich morphological systems. The existing NLP task for Inuktitut is that of word alignment (Martin et al., 2005), where Inuktitut tokens align to entire English clauses. While Langlais et al. (2005) theorises that a morphological analyser could aid in this task, we observed little to no improvement over a baseline model by making use of its segmentation. Nonetheless, morphological analysis does provide a great deal of information, but the task structure tends to disprefer its contribution.

2 Background

2.1 Inuktitut

Inuktitut is a macrolanguage of many more-or-less mutually intelligible dialects (Gordon, 2005). The morphosyntax of Inuktitut is particularly marked by a rich polysynthetic suffixing morphology, including incorporation of arguments into verbal tokens, as in *natsiviniqtulauqsimavilli* in (1). This phenomenon causes an individual token in Inuktitut to be approximately equivalent to an entire clause in English.

(1)	*natsiq-*	*-viniq-*	*-tuq-*	*-lauq-*	*-sima-*
	seal	meat	eat	before	ever
	-vit	*-li*			
	INT-2s	but			

“But have you ever eaten seal meat before?”

Lowe (1996) analyses the morphology as a four-place relationship: one head morpheme, zero or more lexical morphemes, one or more grammatical morphemes, and an optional enclitic. The morphotactics causes, amongst other phenomena, the final consonant of a morpheme to assimilate the manner of the initial consonant of the following morpheme (as in *-villi*), or to be dropped (as in *natsiviniq-*). Consequently, morphemes are not readily accessible from the realised surface form, thereby motivating the use of a morphological analyser.

2.2 Morphological analysis

For many languages with a less rich morphology than Inuktitut, an inflectional lexicon is often adequate for morphological analysis (for example, CELEX for English (Burnage, 1990), Le*fff* for French (Sagot et al., 2006) or Adolphs (2008) for German). Another typical approach is to perform morphological analysis at the same time as POS tagging (as in Hajič and Hladká (1998) for the fusional morphology in Czech), as it is often the case that

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 372–376,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

determining the part-of-speech and choosing the appropriate inflectional paradigm are closely linked.

For highly inflecting languages more generally, morphological analysis is often treated as a segment-and-normalise problem, amenable to analysis by weighted finite state transducer (wFST), for example, Creutz and Lagus (2002) for Finnish.

3 Resources

3.1 A morphological analyser for Inuktitut

The main resource that we are evaluating in this work is a morphological analyser of Inuktitut called Uqa·Ila·Ut.[1] It is a rule-based system based on regular morphological variations of about 3200 head, 350 lexical, and 1500 grammatical morphemes, with heuristics for ranking the various readings. The head and lexical morphemes are collated with glosses in both English and French.

3.2 Word alignment

The training corpus we use in our experiments is a sentence-aligned segment of the Nunavut Hansards (Martin et al., 2003). The corpus consists of about 340K sentences, which comprise about 4.0M English tokens, and 2.2M Inuktitut. The challenge of the morphology becomes apparent when we contrast these figures with the types: about 416K for Inuktitut, but only 27K for English. On average, there are only 5 token instances per Inuktitut type; some 338K types (81%) are singletons.

Inuktitut formed part of one of the shared tasks in the ACL 2005 workshop on building and using parallel texts (Martin et al., 2005); for this, the above corpus was simplistically tokenised, and used as unsupervised training data. 100 sentences from this corpus were phrasally aligned by Inuit annotators. These were then extended into word alignments, where phrasal alignments of one token in both the source and target were (generally) called sure alignments, and one-to-many or many-to-many mappings were extended to their cartesian product, and called probable. The test set was composed of 75 of these sentences (about 2K English tokens, 800 Inuktitut tokens, 293 gold-standard sure alignments, and 1679 probable), which we use to evaluate word alignments.

Our treatment of the alignment problem is most similar to Schafer and Drábek (2005) who examine four systems: GIZA++ models (Och and Ney, 2000) for each source-target direction, another where the Inuktitut input has been syllabised, and a wFST model. They observe that aggregating these results through voting can create a very competitive system for Inuktitut word alignment.

4 Experimental approach

We used an out-of-the-box implementation of the Berkeley Aligner (DeNero and Klein, 2007), a competitive word alignment system, to construct an unsupervised alignment over the 75 test sentences, based on the larger training corpus. The default implementation of the system involves two jointly-trained HMMs (one for each source-target direction) over five iterations,[2] with so-called competitive thresholding in the decoding step; these are more fully described in DeNero and Klein (2007) and Liang et al. (2006).

Our approach examines morphological pre-processing of the Inuktitut training and test sets, with the idea of leveraging the morphological information into a corpus which is more amenable to alignment. The raw corpus appears to be under-segmented, where data sparseness from the many singletons would prevent reliable alignments. Segmentation might aid in this process by making sub-lexical units with semantic overlap transparent to the alignment system, so that types appear to have a greater frequency through the data. Through this, we attempt to examine the hypothesis that one-to-one alignments between English and Inuktitut would hold with the right segmentation. On the other hand, oversegmentation (for example, down to the character level) can leave the resulting sub-lexical items semantically meaningless and cause spurious matches.

We consider five different ways of tackling Inuktitut morphology:

1. **None**: simply treat each Inuktitut token as a monolithic entity. This is our baseline approach.

[1] `http://inuktitutcomputing.ca/Uqailaut/en/IMA.html`

[2] Better performance was observed with three iterations, but we preferred to maintain the default parameters of the system.

2. **Head**: attempt to separate the head morpheme from the non-head periphery. Our hypothesis is that we will be able to align the clausal head more reliably, as it tends to correspond to a single English token more reliably than the other morphemes, which may not be realised in the same manner in English. Head morphs in Inuktitut correspond to the first one or two syllables of a token; we treated them uniformly as two syllables, as other values caused a substantial degredation in performance.

3. **Syllabification**: treat the text as if Inuktitut had isolating morphology, and transform each token into a series of single-syllable pseudo-morphs. This effectively turns the task on its head, from a primarily one Inukitut-to-many English token problem to that of one English-to-many Inuktitut. Despite the overzealousness of this approach (as most Inuktitut morphemes are polysyllabic, and consequently there will be many plausible but spurious matches between tokens that share a syllable but no semantics), Schafer and Drábek (2005) observed it to be quite competitive.

4. **Morphs**: segment each word into morphs, thereby treating the morphology problem as pure segmentation. This uses the top output of the morphological analyser as the oracle segmentation of each Inuktitut token.

5. **Morphemes**: as previous, except include the normalisation of each morph to a morpheme, as provided by the morphological analyser, as a sort of "lemmatisation" step. The major advantage over the morph approach is due to the regular morphophonemic effects in Inuktitut, which cause equivalent morphemes to have different surface realisations.

5 Results

5.1 Analyser

In our analysis, the morphological analyser finds at least one reading for about 218K (= about 65%) of the Inuktitut types. Of the 120K types without readings, resource contraints account for about 11K. [3] Another 6K types caused difficulties due to punctuation, numerical characters or encoding issues, all of which could be handled through more sophisticated tokenisation.

A more interesting cause of gaps for the analyser was typographical errors (e.g. **kiinaujaqtaarua̲sirnirmut* for *kiinaujaqtaarusia̲rnirmut* "requests for proposals"). This was often due to consonant gemination, where it was either missing (e.g. *nunavum̲mut* "in Nunavut" appeared in the corpus as **nunavumut*) or added (e.g. **tamakkunun̲nga* instead of *tamakkununga* "at these ones here"). While one might expect these kinds of error to be rare, because Inuktitut has an orthography that closely reflects pronunciation, they instead are common, which means that the morphological analyser should probably accept incorrect gemination with a lower weighting.

More difficult to analyse directly is the impact of foreign words (particularly names) — these are typically subjectively transliterated based on Inuktitut morphophonology. Schafer and Drábek (2005) use these as motivation for an approach based on a wFST, but found few instances to analyse its accuracy. Finally, there are certainly missing roots, and possibly some missing affixes as well, for example *pirru-* "accident" (cf. *pirruaqi-* "to have an accident"). Finding these automatically remains as future work.

As for tokens, we briefly analysed the 768 tokens in the test set, of which 228 (30%) were not given a reading. Punctuation (typically commas and periods) account for 117 of these, and numbers another 7. Consonant gemination and foreign words cause gaps for at least 16 and 6 tokens, respectively (that we could readily identify).

5.2 Word Alignment

Following Och and Ney (2000), we assess using alignment error rate (AER) and define precision with respect to the probable set, and recall with respect to

[3] We only attempted to parse tokens of 30 characters or shorter; longer tokens tended to cause exceptions — this could presumably be improved with a more efficient analyser. While the number of analyses will continue to grow with the token length, which has implications in agglutinative languages, here there are only about 300 tokens of length greater than 40.

Approach	Prec	Rec	AER
None	0.783	0.863	0.195
Head	0.797	0.922	0.176
Syllabification	0.789	0.881	0.192
Morphs	0.777	0.860	0.207
Morphemes	0.777	0.863	0.206
S&D E-I	0.646	0.829	0.327
S&D Syll	0.849	0.826	0.156

Table 1: Precision, recall, and alignment error rate for various approaches to morphology, with Schafer and Drábek (2005) for comparison

the sure set.

We present word alignment results of the various methods — contrasted with Schafer and Drábek (2005) — in Table 1. The striking result is in terms of statistical significance: according to χ^2, most of the various approaches to morphology fail to give a significantly ($P < 0.05$) different result to the baseline system of using entire tokens. For comparison, whereas our baseline system is significantly better than the baseline system of Schafer and Drábek (2005) — which demonstrates the value that the Berkeley Aligner provides by training in both source-target directions — their syllablised model is significantly superior in precision ($P < 0.001$), while their recall is still worse than our model ($P < 0.05$). Intuitively, this seems to indicate that their model is making fewer judgments, but actually the opposite is true. It seems that their model achieves better performance than ours because it leverages many candidate probable alignments into high quality aggregates using a most-likely heuristic on the mapping of Inuktitut syllables to English words, whereas the Berkeley Aligner culls the candidate set in joint training.

Of the approaches toward morphology that we consider, only the recall of the head–based system improves upon the baseline ($P < 0.025$). This squares with our intuitions, where segmenting the root morpheme from the larger token allows for more effective alignment of the semantically straightforward sure alignments.

The three systems that involve a finer segmentation over the tokens are equivalent in performance to the baseline system. The oversegmentation seemed to caused the alignment system to abandon an implicit preference for monotonicity of the order of tokens between the source and target (which holds pretty well for the baseline system over the test data, thanks partly to the fidelity-focused structure of a Hansard corpus): presumably because the aligner perceives lexical similarity between disparate tokens due to them sharing a sublexical unit. This relaxing of monotonicity is most apparent for punctuation, where a comma with a correct alignment in the baseline becomes incorrectly aligned to a different comma in the sentence for the segmented system.

6 Conclusion

The only improvement toward the task that we observed using morphological approaches is that of head segmentation, where using two syllables as a head-surrogate allowed us to capture more of the sure (one-to-one) alignments in the test set. One possible extension would be to take the head morpheme as given the analyser, rather than the somewhat arbitrary syllabic approach. For other languages with rich morphology, it may be similarly valuable to target substantives for segmentation to improve alignment.

All in all, it appears that the lexical encoding of morphology of Inuktitut is so strikingly different than English, that the assumption of Inuktitut morphemes aligning to English words is untrue or at least unfindable within the current framework. Numerous common morphemes have no English equivalent, for example, *-liaq-* “to go to” which seems to act as a light verb, or *-niq-*, a (re-)nominaliser for abstract nominals. While the output of the morphological analyser could probably be used more effectively in other tasks, there are still important impacts in word alignment and machine translation, including leveraging a dictionary (which is based on morphemes, not tokens, and as such requires segmentation and normalisation) or considering grammatical forms for syntactic approaches.

References

Peter Adolphs. 2008. Acquiring a poor man’s inflectional lexicon for German. In *Proc. of the 6th LREC*, Marrakech, Morocco.

Gavin Burnage. 1990. CELEX: A guide for users. Technical report, University of Nijmegen.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. of the 6th Workshop of ACL SIGPHON*, pages 21–30, Philadelphia, USA.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proc. of the 45th Annual Meeting of the ACL*, pages 17–24, Prague, Czech Republic.

Raymund G. Gordon, Jr, editor. 2005. *Ethnologue: Languages of the World, Fifteenth Edition*. SIL International.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on COLING*, pages 483–490, Montréal, Canada.

Philippe Langlais, Fabrizio Gotti, and Guihong Cao. 2005. NUKTI: English-Inuktitut word alignment system description. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 75–78, Ann Arbor, USA.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proc. of the HLT Conference of the NAACL*, pages 104–111, New York City, USA.

Ronald Lowe. 1996. Grammatical sketches: Inuktitut. In Jacques Maurais, editor, *Quebec's Aboriginal Languages: History, Planning and Development*, pages 204–232. Multilingual Matters.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proc. of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts*, pages 115–118, Edmonton, Canada.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, USA.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the ACL*, pages 440–447, Saarbrücken, Germany.

Benoît Sagot, Lionel Clément, Eric Villemonte de La Clergerie, and Pierre Boullier. 2006. The Le*fff* syntactic lexicon for French: Architecture, acquisition, use. In *Proc. of the 5th LREC*, pages 1348–1351, Genoa, Italy.

Charles Schafer and Elliott Drábek. 2005. Models for Inuktitut-English word alignment. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 79–82, Ann Arbor, USA.