# Symmetrized Word Alignments for English-Inuktitut Machine Translation

Jeff Pinner - SUID 5215534
CS224N Final Project
3/11/11 - Submitted 2 Days Late

**The Inuktitut Language**

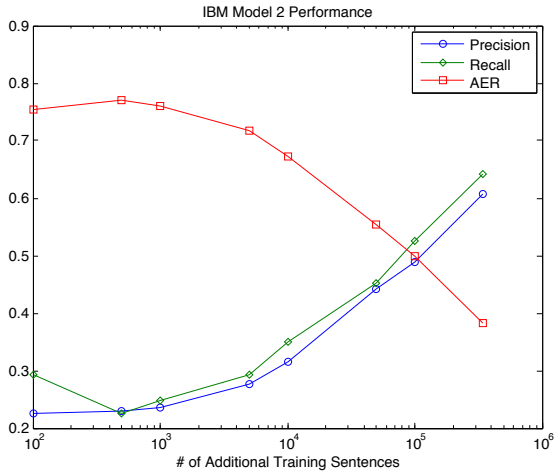
Inuktitut is a highly agglutinative language, with words formed from a combination of many individual morphemes. This is readily evident in the Hansards from the Canadian Legislative Assembly of Nunavut that comprise the ACL 2005 training corpus. English sentences in this corpus are on average almost twice as long as the corresponding Inuktitut sentence. Over 330,000 different Inuktitut words are represented in the corpus as compared with under 9000 different English words. Some examples or Inuktitut to English word translations illustrating this are provided below:

| Inuktitut Word | English Phrase |
|---|---|
| kiinaujanik | money / some funds |
| kiinaujanut | with the funds |
| kiinaujaqaqtipanngitavut | it is not something that we fund |

This creates several difficulties that cause the IBM alignment models to perform poorly. The first has to do with the word alignments themselves. When translating from Inuktitut to English, the IBM models do not allow for many English words to be aligned to a single Inuktitut word, but this many-to-one alignment accounts for the majority of the alignments in the language pair. The second difficulty has to do with translation probabilities. Even assuming perfect alignment knowledge, the abundance of Inuktitut words compared to English words does not allow the model to discriminate between correct and incorrect translations since any single English-Inuktitut word pair occurs infrequently, regardless of the accuracy of the pair translation.

**IBM Alignment Models for Inuktitut to English Translation**

English and Inuktitut have similar sentence order and structure, resulting in alignments between the sentences that tend to be along the diagonal. This feature of the language pair, along with the fact that the common many-to-one alignment between Inuktitut and English words result in most English words having a fertility of one in the IBM alignment models, led to the selection of the IBM Model 2 alignment model for Inuktitut to English translation. The performance of the IBM Model 2 alignment model on the ACL 2005 training corpus is plotted below as a function of the size of the training set.



The improvement of the model is log-linear with training set size when training on 5000 sentence pairs or more. A maximum in the performance curve of the model is not achieved even after training on the entire ACL 2005 training corpus (340,256 sentence pairs). Better performance could still be achieved with a larger training set size due to the sheer number of Inuktitut words. More complex alignment models are not helpful in improving alignment performance with this size corpus. As an example, after training on the entire ACL 2005 training corpus, the IBM Model 2 alignment model achieves precision, recall, and AER scores of around 60%, 64%, and 38%, respectively on the validation data. The IBM Model 3 alignment model achieves scores of around 23% for precision and recall, with an AER of around 77%. The poor performance of the more complex alignment models is partly due to the difference in sentence length between English and Inuktitut sentences. Longer English sentences cause the IBM alignment models to assign high zero-fertility probabilities to words. This can be seen in the comparison of the two alignment models below on an exemplar sentence:

```
IBM Model 2 Alignment:
 #                                                      | taikkua
( )( )( )( )( )(#)( )                                   | pijumanirsaujariaqarmata
                     ( )(#)                             | pijumajaujunik
                            ( )( )( )             #     | nunaliujunik
                                     [#]                | ammalu
                                        (#)( )          | piliriqatiqarlutik
                                               ( )(#)   | nunaliujunik
                                                     [ ]| .
-------------------------------------------------------'
 T  h  t  b  m  r  t  t  n  o  t  c  a  w  w  t  c  .
 h  a  o  e  o  e  o  h  e  f  h  o  n  o  i  h  o
 e  v        r  c     e  e     e  m  d  r  t  e  m
 y  e        e  e        d        m     k  h     m
               p        s        u              u
               t                 n              n
               i                 i              i
               v                 t              t
               e                 i              i
                                 e              e
                                 s              s

IBM Model 3 Alignment:
 #                                                      | taikkua
( )( )( )( )( )( )( )                                   | pijumanirsaujariaqarmata
                     ( )( )                             | pijumajaujunik
                            ( )( )( ) #                 | nunaliujunik
                                     [ ]                | ammalu
                #                       ( )( )          | piliriqatiqarlutik
                                               ( )( )   | nunaliujunik
                                                     [#]| .
-------------------------------------------------------'
 T  h  t  b  m  r  t  t  n  o  t  c  a  w  w  t  c  .
 h  a  o  e  o  e  o  h  e  f  h  o  n  o  i  h  o
 e  v        r  c     e  e     e  m  d  r  t  e  m
 y  e        e  e        d        m     k  h     m
               p        s        u              u
               t                 n              n
               i                 i              i
               v                 t              t
               e                 i              i
                                 e              e
                                 s              s
```

Most English words are represented by morphemes in the Inuktitut words and thus simply aligning the Inuktitut words close to the diagonal results in lower AER scores.

**IBM Alignment Models for English to Inuktitut Translation**

Since Inuktitut is highly agglutinative, and since almost all English words are represented by Inuktitut morphemes, reverse alignments can be used to improve the performance of the IBM alignment models. Consider the example sentence and the forward alignment given below.

```
IBM Model 2 Alignment:
( )(#)( )( )                                          | qaujisariaqaqpavut
            ( )( )( )( )( )( )(#)                     | kiujunnaqtinniqsauniaruptigu
                                          #       [ ]| uplumi
                                            ( )(#)    | nunavumi
                                  # ( )( )( )          | iliqusirijauliqtumut
-------------------------------------------------,
 W  m  r  t  i  o  t  m  t  m  r  t  t  r  o  N  t
 e  u  e  h  n  r  o  a  h  o  e  o  h  e  f  u  o
    s  v  i     d     k  e  r  s     e  a     n  d
    t  i  e     e     e  m  e  p        l     a  a
       e        r              o        i     v  y
       w                       n        t     u
                               s        y     t
                               i
                               v
                               e
```

The English sentence is more than 3 times longer than its corresponding Inuktitut sentence. Given the lack of many-to-one alignments in the IBM models, these models will not be able achieve good performance on the example regardless of the amount of training data available. The reverse alignment however does not suffer from this same restriction. For the example above, the reverse alignment given by the IBM Model 2 alignment model (shown on the following page) does align every English word to an Inuktitut word (albeit not completely correctly). Nevertheless, it provides a good starting point from which we can generate symmetrized alignments to improve performance given the training set size limitations. The IBM Model 2 alignment model out-performs the Model 3 alignment model in the reverse direction as well. In the reverse direction, the Model 3 alignment model overestimates the Inuktitut word fertilities.

```
IBM Model 2 Reverse Alignment:
(#)             | We
( ) #           | must
( ) #           | review
( ) #           | them
   ( )    #     | in
   ( )       #  | order
   ( ) #        | to
   (#)          | make
   (#)          | them
   ( )       #  | more
   ( )       #  | responsive
       #    ( )| to
            ( )| the
            (#)| reality
         (#)    | of
         (#)    | Nunavut
      [#]       | today
---------------'
 q  k  u  n  i
 a  i  p  u  l
 u  u  l  n  i
 j  j  u  a  q
 i  u  m  v  u
 s  n  i  u  s
 a  n     m  i
 r  a     i  r
 i  q        i
 a  t        j
 q  i        a
 a  n        u
 q  n        l
 p  i        i
 a  q        q
 v  s        t
 u  a        u
 t  u        m
    n        u
    i        t
    a
    r
    u
    p
    t
    i
    g
    u
```

```
IBM Model 3 Reverse Alignment:
( ) #           | We
( ) #           | must
( ) #           | review
( )    #        | them
   (#)          | in
   ( )       #  | order
   (#)          | to
   (#)          | make
   ( ) #        | them
   ( )       #  | more
   ( )       #  | responsive
            (#)| to
            (#)| the
            (#)| reality
       # ( )    | of
    #    ( )    | Nunavut
      [ ]    #  | today
---------------'
 q  k  u  n  i
 a  i  p  u  l
 u  u  l  n  i
 j  j  u  a  q
 i  u  m  v  u
 s  n  i  u  s
 a  n     m  i
 r  a     i  r
 i  q        i
 a  t        j
 q  i        a
 a  n        u
 q  n        l
 p  i        i
 a  q        q
 v  s        t
 u  a        u
 t  u        m
    n        u
    i        t
    a
    r
    u
    p
    t
    i
    g
    u
```

**Symmetrized Word Alignments**

Symmetrized alignments were created by training the IBM Model 2 alignment model in both the forward and reverse directions and then combining the alignments using the heuristic method given in [Och, 2004]. A comparison of the performance of the symmetrized alignment model and the IBM Model 2 alignment models on the ACL 2005 validation and test data sets is summarized in the table below:

| Alignment Model | Validation Data Set Precision | Validation Data Set Recall | Validation Data Set AER |
|---|---|---|---|
| IBM Model 2 | 0.606 | 0.642 | 0.382 |
| Symmetrized Model | 0.693 | 0.832 | 0.275 |
| | Test Data Set Precision | Test Data Set Recall | Test Data Set AER |
| IBM Model 2 | 0.520 | 0.512 | 0.482 |
| Symmetrized Model | 0.531 | 0.819 | 0.430 |

**Future Work**

Given the agglutinative nature of the Inuktitut language, further improvement might be achievable if instead of trying to find word-to-word alignments, the alignment models were trained on morpheme-to-word alignments. Dividing the Inuktitut words into their constituent morphemes would help alleviate (or possible eliminate entirely) the need for many-to-one alignments and the inaccuracies in the fertility probabilities of the more complex alignment models.

**References**

Brown, Peter F., et. al. “The Mathematics of Statistical Machine Translation,” Computational Linguistics, Vol 19, No 2, pp. 263 - 311, 1993.

Och, Franz Josef and Ney, Hermann, “The Alignment Template Approach to Statistical Machine Translation,” Computational Linguistics, Vol 30, No 4, pp. 417 - 449, 2004.