university of groningen / faculty of science and engineering

An Evaluation of Parallel Text Extraction and Sentence Alignment for Low-Resource Polysynthetic Languages

Bachelor's Project Thesis

Kevin Kelly (k.kelly.1@student.rug.nl)
Supervisors: dr. J.K. Spenader

Abstract: For the development of robust NLP applications such as building accurate machine translation systems, large monolingual and parallel corpora are essential. Many polysynthetic languages, where words are built up out of several concatenated morphemes, lack such resources. Their high morpheme to word ratio and sometimes complex morphological structure make creation of these resources problematic. This research explores if high quality alignment between a polysynthetic language and non-polysynthetic language is possible when using existing sentence alignment tools, and if tools to automatically harvest bitexts from multilingual sites can be used to produce large amounts of meaningful parallel data between these languages. Alignment between Inuktitut and English, and Kalaallisut and Danish was evaluated. In an Intrinsic evaluation, a method to obtain accuracy, recall and F1 scores in absence of a gold standard through the use of tf-idf and co-occurence statistics was designed and evaluated. Results show high F1 scores when used with a one co-occurence "word pair" per aligned sentence threshold on segmented polysynthetic data, but degrades when higher threshold limits are set. In an extrinsic evaluation, a neural machine translation model showed improved CHRF scores for Inuktitut to English translation when 1134 aligned sentences were added to an existing training set. Poor translation quality was however shown between Danish and Kalaallisut when trained solely on 14778 aligned sentences acquired through parallel text extraction from multilingual websites.

1 Introduction

In computational linguistics, the use of a corpora is one of the cornerstones for successful natural language processing. For a corpus to be able to provide a useful basis for any further NLP processes it is important that the compilation of the corpora is reliable and of good quality. For tasks such as developing machine translation systems, these conditions extend to the use of parallel corpora where the added prerequisite of alignment between documents, paragraphs, sentences and in some cases words between the source language and target language of the corpora are of high consequence in the development of well performing translation systems (Tiedemann, 2011).

For many modern MT systems parallel texts come from sources such as parliamentary proceedings of multinational institutions, including the "Europarl Corpus" (Koehn, 2005), a corpus aligned at the sentence level containing 30 million words from the 11 official languages of the European union, and the "MultiUN" (Tiedemann, 2012), a corpus created from United Nations documentation containing 300 million words for each of the 6 languages of the United nations. Other useful parallel resources come from government documentation of multilingual countries such as the "The Canadian Hansard Corpus" (Beelen et al., 2017) a corpus containing 1.3 million pairs of aligned text chunks from official records in French and English. While these types of corpora continue to be a useful source for the improvement of MT systems they are limited in their range of languages. The majority of languages in the world lack such parallel data making the development of robust MT translation systems between a large amount of languages unfeasible.

The goal of this project has been to look into methods of gathering parallel texts and evaluating automatic sentence alignment between parallel texts for low-resource polysynthetic languages. It focuses on two languages of the Eskimo-Aleut language family. That of Inuktitut, one of the principal Inuit languages of Canada, and Kalaallisut, the main language spoken in Greenland, and looks at the process of sentence alignment between Inuktitut and English as well as Kalaallisut and Danish.
For purposes of the evaluation on alignment quality, a method using tf-idf and co-occurrence statistics is developed, as no gold standard or manually aligned and annotated corpora existed for these languages at the time of starting this project. The method proposed is also evaluated on two other language pairs with differing levels of morphological complexity in order to establish the robustness of evaluation. For this purpose, English and Swedish as well as English and Finnish were selected. English and Swedish both contain very little morphology whereas Finnish contains a lot of inflectional morphology but is not a polysynthetic language. The two language pairs are used to compare alignment quality on languages with different levels of morphological complexity
In this paper I look to answer if alignment on the sentence level is achievable for low-resource polysynthetic languages, as well as explore the difficulties that these languages present for sentence alignment. The project will also look to answer if a novel method of evaluation is a plausible metric in comparison to the traditional use of manually aligned gold standard in determining accuracy, recall and F1 scores.
Section 2 provides a brief background in the use of sentence alignment tools as well as the nature of polysynthetic languages, and challenges posed by them in regards to natural language processing.
Section 3 will discuss the tools used for this project and section 4 the pre-processing steps taken to ensure efficiency of alignment and parallel text retrieval.
In section 5 the paper will evaluate the results of alignment based on the novel method introduced in this paper as well as evaluate alignment and parallel text retrieval through the use of neural machine translation.

2 Background

The alignment of polysynthetic languages follows the same key steps used in the construction of any parallel texts. The process is done hierarchically starting at the more general level of alignment and increasing in specificity of alignment at each subsequent level. Each level of alignment requires the ith segment in the source text to have its corresponding translation as the ith segment in the target text (Tiedemann, 2011).

1. Document alignment - Commonly the first step of alignment between large amounts of texts is aligning corresponding documents between languages. This allows subsequent levels of alignment to be done appropriately as well as providing easy look up between texts. For this purpose an appropriate mapping between documents needs to be done, and depending on the amount of documents and method used to collect the data this is not always a straight forward task, especially when web based methods of data collection are used. Issues such as inconsistent filenames can make identifying which web-pages are translations of each other difficult, and need to be remedied at this level.

2. Paragraph alignment - Once the documents are aligned the contents need to be aligned on the paragraph level. Paragraph alignment is comparatively easier than the other levels of alignment. When done automatically paragraph alignment commonly relies on length based metrics such as number of sentences. Most alignment tools that exist for this stage of alignment are robust and accurate. Issues occur more frequently in the case of literary texts. Many tools used for sentence alignment include methods to handle paragraph alignment as well

3. Sentence alignment - Sentence alignment is the process of aligning the texts so that sentences of the source language and its corresponding sentence translation are found in the same place in both texts. This level of alignment is often a more complex and difficult task than the above levels as the increased specificity of alignment results in more complications. Issues of alignment include sentences missing from

one text that are present in the other text, a sentence in one text corresponding to two or more sentences in the other text. The order of sentences can also be different in the two texts as shown in figure 2.1
When done manually in compilation of corpora this is one of the more labour intensive and time consuming processes (Tiedemann, 2011). The above three levels of alignment rely to a certain extent on a monotonic constraint which entails that there are no crossing links between segments (see figure 2.1). This would imply on the sentence level that sentence numbers will not commonly cross each other in the alignment process, i.e. 1-2 and 2-1 will very rarely happen together. As reordering of sentences are not common when translating texts this constraint often holds and by applying this constraint the search space is largely reduced and simpler methods of alignment can be applied. In more modern alignment tools however this is treated as a soft constraint, where merging of sentences to create monotonic alignment will be applied in cases of seemingly non-monotonic sentence equivalency.
Document alignment was for most of the gathered material done manually, and so the focus of this project has mainly been on the level of sentence alignment.

2.1 Different types of aligners

Many NLP applications such as machine translation focus on the use of aligned sentences as they provide a source that is localized enough to provide important grammatical, semantic and lexical information but not general or large enough to introduce too much noise to the data. Due to the importance of parallel sentences in NLP systems, several sentence alignment tools have been developed over the years. Some aligners are more general and language independent, while other aligners have been developed specifically to work between certain languages or for specific cases, such as noisy data or where non-monotonic sentence equivalency is very prevalent (Quan et al., 2013). Generally speaking sentence alignment tools can be divided into three categories (Seničić and Fairon, 2017). Each type is presented below with details about one example system specific to that type of alignment
Statistical aligners use length based algorithms and are entirely statistical in nature, and look at sentence length to determine plausible equivalency of sentences.

- Gale-Church – The Gale-Church algorithm (Gale and Church, 1993) is an example of a statistical aligner. It is an unsupervised sentence alignment algorithm that works on the idea that similar sentences will often be of similar length. It uses character length of sentences rather than number of words to assign similarity scores. In the original paper it was used on a sub-corpus consisting of 725 sentences from the UBS English-French, and showed an alignment accuracy of 96% . The algorithm relies on the monotonic constraint that there are no crossing links between sentences. Despite its simplicity, length measurements have been shown to be enough for high alignment accuracy between many languages under the assumption that the same information is displayed in both texts. After its initial release it has subsequently been used in the alignment of the Europarl corpus (Koehn, 2005) where by using meta-textual information such as timestamps, speaker information, paragraph and chapter separation to provide anchors, alignment quality was judged to be very high. Most modern Hybrid aligners rely at least partially on the Gale-Church algorithm for statistical based alignment.

Lexical aligners use lexical information such as dictionaries or lexicons to determine a relationship between sentences in the source and target language.

- Champollion - Champollion (Ma, 2006) is a lexicon based sentence aligner. It was initially developed for the alignment of English - Chinese parallel texts but was later ported to work on other languages. Champollion assumes that the parallel text data my be noisy i.e. that a large percentage of potential alignments will not be 1-1 alignments and that the number of insertions and deletions will be significant. In order to improve alignment quality it relies on a “Term Frequency - Inverse Document Frequency” metric which allows the system to

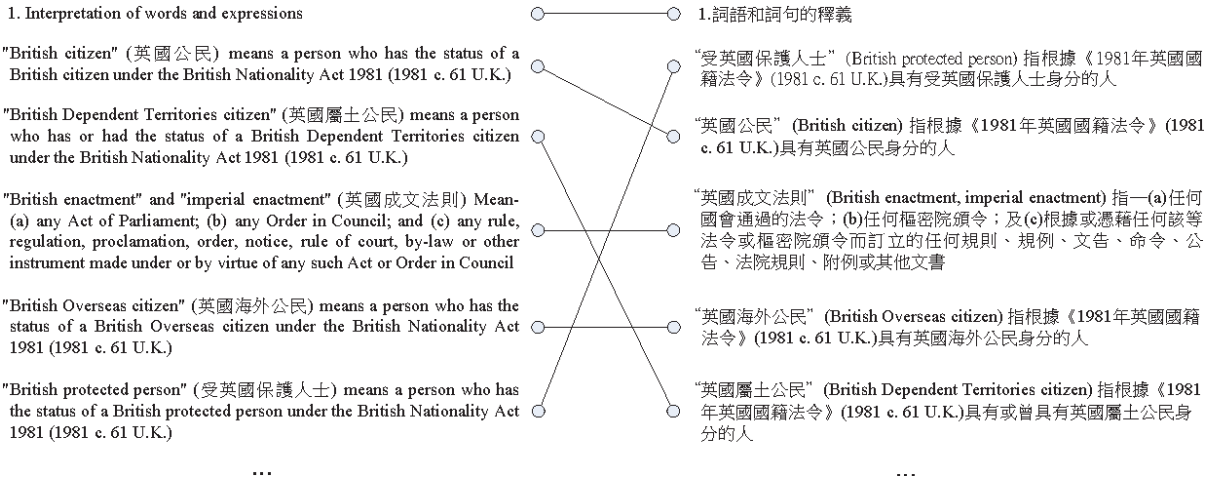

Figure 2.1: Example of non-monotonic sentence alignment from BLIS corpus, taken from (Quan et al., 2013)

assign greater weight to less frequently translated words, and uses sentence length information to weed out alignments that seem improbable. Champollion achieved 97.0% precision and 96.9% recall. for 1-1 sentences and 54.6% precision and 45.3% recall for 1-0 and 0-1 sentences, when using a 58000 word lexicon on a test set consisting of 3788 English sentences and 3866 Chinese sentences, and achieved an overall precision of 97.0% and 96.0% recall (Ma, 2006).

Hybrid aligners combine the use of statistical methods with available lexical information.

- Hunalign - Hunalign (Varga et al., 2007) is a hybrid approach to sentence alignment. It was developed to be used on medium density languages, encompassing languages between those which have very high availability of digital materials and those which have very low availability of digital materials. It can be used for all language pairs but was developed for English - Hungarian. It uses statistical length measurements based on the Gale-Church algorithm coupled with any lexicon information provided for the languages.
  Hunalign can be specified to use a bootstrap method to add words to the lexicon, or used when no lexicon is present. When used on George Orwells “1984” (Varga et al., 2007) with length + dictionary, Hunalign achieved 99.34% precision and 99.34% recall. When used with length + bootstrap method but without a dictionary it achieved 99.12% precision and 99.18% recall.
  Hunalign has been shown to work on languages characterized by high degrees of inflection such as Polish (Wołk and Marasek, 2014) where Hunalign was one of the top performing aligners investigated. The authors designed their own method of evaluation based on correct alignment pairs and miss-aligned pairs. The score given was then normalised between 1 and 100 with Hunalign having a 97.85 score.
  Hunalign was also used to align the JRC-Acquis corpus (Steinberger et al., 2006), a parallel corpus of 21 European languages with an average of 8.8 million tokens in 7,600 texts per language which contains highly inflected languages such as Finnish and Polish. As it is a cross lingual corpora it is however hard to determine the overall quality across languages.

.

2.2 Challenges of low resource polysynthetic languages in sentence alignment

The definition of what constitutes a polysynthetic language is up for some debate. Many researchers have offered variations of what the term entails as

well as what should be needed for a language to constitute a polysynthetic one (Mithun, 2009).

Broadly speaking, languages can be classified into different groups based on similarities of their morphological structure. On one end of the spectrum, languages belong to the classification of analytic languages where words have a very low morpheme to word ratio. On the other side of the spectrum languages are classified as synthetic in which syntactic relations within sentences are expressed by inflection or by agglutination. Synthetic languages tend to have a high morpheme to word ratio, where words are created by affixing morphemes to a root. Synthetic languages can further be subdivided into classifications of agglutinative, fusional and polysynthetic.
Agglutinative languages have a high morpheme to word ratio but the morphemes are very distinct having only one meaning for each morpheme, meaning that segmentation can easily be used to divide a word into its meaningful units. See example (1) of an agglutinative word in turkish

(1) evlerinizden
ev-ler-iniz-den
house-plural-your-from
from your house

Fusional languages predominantly use inflectional morphemes, where morphemes often belong to more than one word class. Unlike agglutinantive languages there is no clear cut between morphemes, instead forms of the words themselves change to indicate how they relate to the other words in a sentence. These languages are thus more difficult to segment in a meaningful way.
Polysynthetic languages are not limited to agglutinative or fusional. Depending on the language, they can exhibit varying degrees of both affixation and fusion of morphemes. Additionally certain polysynthetic languages implement noun incorporation, incorporating the subject and object nouns to form complex verb forms which causes words with multiple stems to occur in a single word. Polysynthetic languages have no defined upper limit to how many times morpheme affixation can occur. This leads to very long and very complex words, where a single polysynthetic word can convey the meaning of an entire sentence. Consider the following example from Inuktitut

(2) tusaatsiarunnanngittualuujunga
tusaa-tsia-runna-nngit-tu-alu-u-junga
hear-well-be.able-NEG-DOER-very-BE-PART.1.S
I can't hear very well

A study by J.Greenberg (Greenberg, 1960, as cited by Mithun, 2009) showed that Kalaallisut was one of the highest morpheme per word languages with an average 3.72 morphemes per word, and work done by Roest (2020) estimated that Inuktitut had an average of 4.39 morphemes per word.
The complex nature of polysynthetic languages has meant that very little properly annotated corpora is available for these languages. In addition to this, issues such as complex verbal morphology, no fixed word order, lack of clear lexical division between nouns and verbs, ambiguity in separation between affixes and clitics, and difficulties of proper identification of root words, means that highly polysynthetic languages become very challenging when attempting to create corpora, as proper tokenization, lemmitization and POS tagging can not be done by traditional means.
The complexity of morphology also causes alignment difficulties in evaluating similarity of sentences when using techniques which are more commonly used between languages that are syntactically and derivationally similar to each other. Similarity measurements based on equivalencies between number of words in sentences or word to word length comparisons lose the desired functionality. Rule based approaches which use lexical information such as dictionaries also become difficult to implement as words repeat on a much less frequent basis due to changes in derivational affixation, see Table 2.1
A common practice to use for agglutinating languages is morphological segmentation. By splitting words into their individual morphemes methods can then be used to asses the individual morphemes. However with languages such as Kalaallisut and Inuktitut which incorporates fusion and noun incorporation this remains problematic as even segmented morphemes will not likely be representative of what the word is supposed to represent.

Table 2.1: Change in Kalaallisut word based on change in transitive and intransitive form of the word "answer" in Danish, taken from (Kristensen, 2010)

intransitivt indikativ	akivoq	han/hun/den/det svarer
transitivt indikativ	akivaa	han/hun/den/det svarer ham/hende/den/det
intransitivt interrogativ	akiva?	svarede han/hun/den/det?
transitivt interrogativ	akiviuk?	svarede du ham/hende/den/det?
intransitivt imperativ	akigit!	svar!
transitivt imperativ	akinnga!	svar mig!
intransitivt optativ	akili	lad ham/hende/den/det svare
transitivt optativ	akilinga	lad ham/hende/den/det svare mig

3 Sentence Aligners and Morphological Segmentation

This project makes use of the Hunalign sentence aligner (Varga et al., 2007). The decision to use Hunalign instead of other alignment algorithms was made due to a few different factors. The nature of polysynthetic languages makes Hunaligns use of the Gale Church algorithm valuable, as a sentence containing roughly the same amount of characters is more likely than it containing the same amount of words. Hunalign has also been shown to work well for more inflectional languages (see section 1.2) and its lack of requirement for any additional information about the languages, made it seem the best candidate for use.
In an attempt to improve accuracy of alignment, morphological segmentation was performed on the parallel texts prior to alignment (see section 3.3), and evaluated alongside the same parallel texts where no morphological segmentation had been performed.
For purposes of automatic text retrieval an application called Bitextor (Espla-Gomis and Forcada, 2010) was used (see section 3.2).

3.1 Hunalign

Hunaligns algorithm [1] works by creating a crude word to word translation of the source text by replacing each token with that of the dictionary translation that has the highest frequency in the target corpus. If a word is not available in the dictionary it will use the word from the source text. The generated translation is then compared to the target text on a sentence to sentence basis. A similarity score will be generated based on token and length similarity. Using a matrix of sentence to sentence alignments between the target text and source text, the score for every sentence pair around the diagonal of the alignment matrix is calculated. This method assumes that sentences far away from the sentence in the source language are not likely to be the equivalent sentence in the target language. The algorithm then looks for the optimal path through the alignment matrix by means of dynamic programming. The algorithm only looks at 1:1, 1:2 and 2:1 sentence matches. After the optimal path has been found, the algorithm will iteratively merge neighboring pairs that are of a 1:2, 2:1, 0:1 or 1:0 sentence match based on the length information of the sentences.
If the algorithm is run without a dictionary it will run in two more stages. The second stage will create an artificial dictionary based on co-occurences of words found in the first stage of the algorithm, which are added to the dictionary if the co-occurence of the words are above a certain threshold. It will then repeat the first step using this dictionary. The resulting aligned sentences are then generated side by side with a confidence value for each alignment based on length similarity and dictionary lookup similarity, see Figure 3.1.

3.2 Bitextor

Bitextor [2] is a free online source application for collecting parallel texts from multi-lingual sites and aligning them using Hunalign. It downloads all the HTML files in a website, pre-processes them into a coherent format and, finally, applies a set of heuristics such as text language comparison, filename extension comparison, file size ratio and total text length difference to select pairs of files which are marked as candidates that contain the same text in two different languages.
As low resource languages often lack the data necessary to be used for translation models this can be used as a stepping stone in the creation of a bilingual corpus for such languages.
Since its original conception newer versions have been released that improves upon the original in terms of pre-processing and post-processing. Bitextor comes with a range of options that can be de-

[1] https://github.com/danielvarga/hunalign

[2] https://github.com/bitextor/bitextor

Taanna tugliriliqluniuk Inuit Nunarjuarmi Katimajingit (ICC) Katimavigiuarn	This is the second time that the ICC General Assembly has been held in	0.166154
Inuvialuit nunalingat quviasugutiqaqtittitilaurivut 20-nut ukiunut nunalirinir	The Inuvialuit community also hosted the event over 20 years ago in 19	0.372043
"Inuvialuit nunangat piullariktuunasugigattigu.	"We consider the Inuvialuit land lucky.	0.210811
1992-ngutillugu, tavvani ilagijauqatau lilauq simagatta Inuit Nunarjuarmi I	In 1992, it was here that we joined the Inuit Circumpolar Council.	0.128571
Katinngagatigisimalirlutigullu 22-nik ukiunik," uqarluni Tatiana Achirgina,	We have been together for 22 years now," said Tatiana Achirgina, Vice C	0.679412
<p>	<p>	0

Figure 3.1: Sample output of Hunalign sentence alignment between English Inuktitut from Inuktitut magazine issue 116

fined in a configuration file. Each option allows for certain customization of how bitextor will run. The options required to run it at its most basic level are the following

1. Language ids following ISO 639-1 syntax for each of the languages used
2. A list of sites for Bitextor to be run on
3. The type of crawler to be used and a source of bilingual information between the two languages such as a bilingual lexicon, an MT system or a parallel corpus.

It has proven useful for corpus creation between English-Croatian (Toral et al., 2017) where Croatian is an under resourced language, as well as for Russian-Kazakh (Zhandos, Aigerim, and Diana, 2017) where Kazakh is an under resourced agglutinative language.
For the English-Croatian data, 23 tourism websites were crawled and produced 64,489 sentence pairs when using an accuracy oriented setting "Bitextor 10-best", whilst producing 48,234 sentence pairs when using a recall oriented setting "Bitextor 1-best". For the Russian-Kazakh data 10 websites were crawled resulting in 5925 sentence pairs. It was not specified in the paper what settings were used.

3.3 Morphological Segmentation

Due to the polysynthetic nature of Kalaallisut and Inuktitut, lexicon based methods as well as Hunaligns built-in method of co-occurence measurements lose a lot of their potential benefits when performing sentence alignment. The same word will appear much less frequently and thus be reliably matched with fewer words in the source language. Morphological segmentation was used as a method to alleviate this issue. The purpose of morphological segmentation is to decompose words into its individual morphemes, through the use of trained segmentation models on the language. See example (3) of the Inuktitut word "akiraqtuqtut"

(3) **Source**: akiraqtuqtut
Target: akiraq @ tuq @ tut

Two segmentation models were used for this purpose. A semi supervised segmentation model "Morfessor FlatCat" was used for Kalaallisut (Mol, 2020). FlatCat proved the most accurate out of the different segmentation models evaluated for the Kalaallisut language, when compared to a validation set.
A combination of a neural based segmentation method, a rule-based segmentation method and a "BPE5K" model (Roest, 2020) was used for the Inuktitut data.

4 Data and Preprocessing

Hunalign was used on English-Finnish, English-Swedish, English-Inuktitut as well as Danish - Kalaallisut. For English - Inuktitut the data was provided by a collection of Inuktitut magazines. [3] Each magazine contained translations in Inuktitut, Inuktitut romanized, English and French. For the purposes of alignment only the texts in Inuktitut romanized and English were used. Extraction of the texts was done manually as the pdf format and placement of texts in the magazine made it difficult for good automatic extraction.
For Danish - Kalaallisut a collection of magazine

[3] https://www.itk.ca/category/inuktitut-magazine/ - issues 115, 116, 119, 120, 121, 122, 125, 126

articles[4] were used that contained translations of both languages for each article. Here also, the extraction of text was done manually. Existing raw corpora for English – Finnish[5] was used to evaluate any contrast in the performance of Hunalign when used on languages with different morphological typology. English – Swedish corpora [6] was also used. Partially as comparison with the other languages but also being a native speaker of both languages it was used as a base point to evaluate Hunaligns performance as well as to judge performance of the novel evaluation method. Both the English-Finnish corpora and the English Swedish corpora were taken from the europarl data repository. Information about all data collected can be found in Table 4.1
The nature of the extracted texts from the magazines meant that some preprocessing was required to make Hunalign work more effectively. Customs scripts were made for this and involved steps to make the data suitable for alignment. The steps taken were:

1. Combine all relevant documents into one text file as Hunalign works better with more data.
2. Remove all end of line hyphen connections to make sure words were not split up.
3. Concatenate texts by removing all unnecessary line breaks to convert texts to normal .txt document format rather than the original article format
4. Add document boundaries using hunaligns "<p>" paragraph marker
5. Perform sentence splitting so that each new sentence is seperated by a newline

For the last stage, an NLTK English sentence splitter was used. The sentence splitter was used on the English, Greenlandic and Inuktitut texts. Even though it was an English sentence splitter it worked well on the Greenlandic and Inuktitut data as the languages share similar sentence boundary disambiguation.

[4]https://timarit.is magazine name - Atuagagdliutit, year 1999, issues 1,2,3,5,6,7

[5]https://www.statmt.org/europarl/ source release, ep-00-01-17, ep-00-01-18, ep-00-02-02, ep-00-02-03, ep-00-02-14

[6]https://www.statmt.org/europarl/ source release, ep-00-01-17, ep-00-01-18, ep-00-02-02, ep-00-02-03, ep-00-02-14

Bitextor was run on 12 different sites containing Danish and Kalaallisut texts, with the Httrack web crawler. Word to word translations were taken from an existing dictionary between Danish and Kalaallisut [7] and provided as a bilingual lexicon.

5 Evaluation

Common evaluation processes for alignment quality normally rely on the use of manually aligned and annotated corpora to use as a comparative gold standard. At the time of making this project no such gold standard existed for Inuktitut-English or Danish-Kalaalisut. Instead a method of "term frequency – inverse document frequency" (tf-idf) was used as a way to establish anchor words. The anchor words were then used to measure co-occurrences of words between sentences to infer if the sentence pairs produced by Hunalign were properly aligned. This method was also used on articles from the europarl corpus in Swedish - English and Finnish – English, where the method could be used on a gold standard to measure if alignment quality was different from the aligned texts produced by Hunalign.

5.1 Finding Anchor Words

Tf-idf is a method to give a numerical statistic that reflects how important a word is to a document in a corpus.

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

$$tf(t, D) = log(1 + freq(t, d))$$

$$idf(t, D) = log(\frac{N}{count(d \in D : t \in d)})$$

Where $t = unique\ term$: $d = document$: $D = set\ of\ all\ documents$

Tf-idf provides a list of weights between 1 and 0 of all unique words in all documents. The higher the term frequency and the lower the document frequency, the more weighted a word is. This results in words that appear frequently throughout all documents to be weighted less heavily, e.g. words such as "the" have a high term frequency but receive a low score as it appears

[7]http://www.ilinniusiorfik.gl/oqaatsit/daka

Table 4.1: Quantity information of language sources

Language	Number of Sentences	Number of Words
English (Inuktitut-magazine)	3483	66023
Inuktitut (Inuktitut-magazine)	3418	37932
Danish (Atuagagdliutit)	2019	23599
Kalaallisut (Atuagagdliutit)	2042	12514
Danish (Bitextor)	160260	756987
Kalaallisut (Bitextor)	160260	570770
English (Europarl)	8994	219927
Swedish (Europarl)	9432	150906
Finnish (Europarl)	8911	150906

frequently throughout all documents.
All articles from the magazines were separated into distinct documents for all 7 languages, with English being done twice for the texts from the Europarl proceedings as well as the texts from the Inuktitut magazines.
"R" was used to calculate the tf-idf score for all words per language. The top 10 words for each document of a language were copied into text files, resulting in 8 word lists separated by language. A few additional words that had a good spread over several different texts but did not have to high a tf-idf score were also included in the word lists for all languages.

5.2 Co-occurrence statistics

Co-occurrence statistics is way to measure if there is an above-chance frequency of items appearing together in some meaningful way. In linguistics it is traditionally used when looking at collocations in a monolingual corpus to measure if there is an above-chance frequent occurrence of two terms alongside each other in a certain order of the corpus. The method has also been used in machine translation evaluation (Doddington, 2002) and word alignment evaluation for STM models (Mi et al., 2014).
In this project co-occurrence measures will be performed on the aligned texts produced by Hunalign to find the most likely word translations for words from the word lists generated by tf-idf. The co-occurrence measure between words was performed in the following way

1. Count number of aligned sentences in parallel texts to get the total number of aligned sentences
2. Separate aligned text into sentences of language1 (lang1) and sentences of language2 (lang2)
3. For each word in the tf-idf word list (wlist1) go through all sentences in lang1. If the word appears in the sentence, save all words in corresponding sentence of lang2 as potential translations to a new list (wlist2).
4. For each word in wlist2 count all occurrences where word1 in lang1 and word2 in lang2 have the same index, and count all occurrences where word2 appears at a sentence index in lang2 but word1 does not appear at the corresponding sentence index of lang1, as well as when word1 appears at an index in lang1 but word2 does not appear in the correspnding index of lang2.

Using the English word "run" and the Swedish translation "spring" we can from the above algorithm acquire the information in table 5.1.

Table 5.1: Co-occurrence measurement between "run" and "spring"

Observed	run	$\neg run$	TOTAL
spring	a	b	a+b
$\neg spring$	c	d	c+d
TOTAL	a+c	b+d	a+b+c+d

The χ^2 value can subsequently be calculated for

each translation candidate in the following way.

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$$

For all collected words in lang2 the word with the highest χ^2 value was then selected and treated as the translation to the word from the source language. This was repeated for all words in the tf-idf list of the source language. The method was then also carried out with the tf-idf list from the target language.
Doing this on all four language pairs gave 8 different lexicons. During the above process the frequency of co-occurrences between a word and its selected translation were recorded. If the frequency of co-occurrences between the selected translation pair was below 4 they were deleted from the lexicon. Any translation duplicates between the two lexicons were also deleted. At this step in the process it was observed that the translations remaining after deletion resulted in smaller lexicons for lexicons that contained polysynthetic languages. It was opted to instead increase the amount of words taken from the tf-idf word lists for all these language pairs to 15 words, in order to produce similar sized lexicons for all languages.

5.3 Evaluating Alignment

The index of an aligned sentence pair was counted for each translation from the lexicon found in the sentence pair. The number of sentences where 1 or more translations were found, 2 or more translations were found, and 3 or more translations were found, were all counted separately and used for the results.
By treating the aligned text as a gold standard and using the co-occurrence results as a measure of if proper alignment occurred or not, precision could be calculated by dividing all sentences containing either 1, 2 or 3 translations by all aligned sentences. For recall, the number of sentences in the unaligned texts was used as an indicator for maximum possible alignments between the two texts. The number of alignments that actually occurred (after accounting for merged sentences) was then divided by the maximum alignments possible, as a way to measure recall.
This method for calculating precision and recall was used for all four language pairs.

5.4 Europarl Corpus Alignment Results

The results for English-Swedish and English-Finnish alignment, can be found in Table 5.2 and Table 5.3 respectively.
For the evaluation of English-Swedish and English-Finnish two evaluations were made. One evaluation focuses on how good Hunaligns alignment quality is between the parallel texts. This was done using the precision and recall methods explained in section 5.2. and shown by the "Null-Dic" columns in table 5.2 and 5.3.
The other evaluation focuses on how well the proposed method of alignment actually works. This is measured by using the evaluation method on a gold standard and calculating the precision. This is shown by the "Gold Standard" columns in table 5.2 and 5.3.
The gold standard used was of the same corpora and was of a similar length to the texts aligned by Hunalign. However, it was found that many of the raw text paragraphs had been deleted in the gold standard. As such, the evaluation on the gold standard corpus does not cover the exact same material as the automatically aligned texts.
As no alignment was performed, only precision is measured on the the gold standard.

Lang	Method	Precision(%)	Recall(%)	F1-score
EN-SW 1 word	Null-Dic	92.0	96.6	94.2
EN-SW 2 word	Null-Dic	78.5	96.6	86.6
EN-SW 3 word	Null-Dic	62.2	96.6	75.7
EN-SW 1 word	Gold Standard	93.1		
EN-SW 2 word	Gold Standard	79.8		
EN-SW 3 word	Gold Standard	63.6		

Table 5.2: Results for English-Swedish parallel texts measured by 1, 2 and 3 occurrences of translated words in a sentence

Lang	Method	Precision(%)	Recall(%)	F1-score
EN-FI 1 word	Null-Dic	93.3	97.0	95.1
EN-FI 2 word	Null-Dic	80.0	97.0	87.9
EN-FI 3 word	Null-Dic	66.2	97.0	78.7
EN-FI 1 word	Gold Standard	93.9		
EN-FI 2 word	Gold Standard	82.1		
EN-FI 3 word	Gold Standard	64.4		

Table 5.3: Results for English-Finnish parallel texts measured by 1, 2 and 3 occurrences of translated words in a sentence

5.4.1 Europarl Corpus Alignment Discussion

The results suggest that alignment quality produced by Hunalign is reasonably good, but only when performed with a "1 translation pair" threshold per sentence. After this performance evaluation degrades significantly the higher the threshold limit is set.
Similar performance rate is measured on the gold standard with only a negligible difference in results from Hunaligns aligned texts. As results should be 100% for the gold standard barring any miss-alignment occurring in the manually aligned corpus, the error rate indicates that the evaluation isn't entirely accurate. As the evaluation works by using translations as anchor words, it is highly dependant on finding a translation in each sentence. The failure to find a translation in each sentence is likely due to some sentences simply not containing any of the listed translation pairs, rather than the sentences being miss-aligned. Using more words from tf-idf lists could be an option to improve results. Another option would be to use more frequently occurring words to increase the spread of words across sentences. However the use of tf-idf was specifically chosen to avoid this, due to the assumption that frequently occurring words are more likely to occur even in miss-aligned sentences, thus resulting in skewed results.
It should be pointed out, that even though the quality of alignment shown is questionable, the use of co-occurrence statistics worked very well in finding good translations when evaluated on translations that occurred together more than three times between Swedish and English. These translations were not always perfect, as partial translations of compound words did occur, such as säkerhetsrådgivare (safety advisers) being translated as advisers. Given that the translations have to appear together in a sentence pair, partial translations were however enough to indicate proper alignment.
The method also served as a good way to find translations appropriate given the context of the texts, but that would not commonly be found in a dictionary. e.g. the indicated translation of president (president in Swedish) was found to be talman (speaker in English), which was the correct translation in the context of the Europarl corpus. Future work combing co-occurrence statistics with a dictionary approach could serve to improve similarity measurements between texts.
It was expected that alignment scores between English-Swedish would be better than that of English-Finnish, due to Finnish being a more inflectional language. This is not what the results show. The higher scores achieved for English-Finnish could be a peculiarity in the use of tf-idf. As Swedish and English are not very morphologically complex languages, common words occur more frequently throughout the text which leads to obscure words scoring higher when using tf-idf, resulting in a list of words that are less representative of the texts. Finnish has more variations of the same word resulting in less over-saturation of more common words in the texts. An example of this is the word "court" in English and "domstol" in Swedish, both appearing very frequently throughout the europarl documents and did not appear in the final lexicon. However in Finnish the word "court" had 3 different translations "tuomioistuimen" "tuomioistuimessa" "tuomioistuin" which were all included in the final lexicon.

5.5 Polysynthetic Alignment Results

The results for English-Inuktitut and Danish-Kalaallisut alignment, can be found in Table 5.4 and Table 5.5 respectively.
For English-Inuktitut and Danish-Kalaallisut evaluation was made on unaltered polysynthetic texts, shown by the "Null-Dic" columns in table 5.4 and 5.5, as well as on segmented polysynthetic texts, shown by the "Segmented" columns of table 5.4 and 5.5. Both evaluations were done using the precision and recall methods explained in section 5.2.

5.5.1 Polysynthetic Alignment Discussion

When used on unaltered polysynthetic sentences, the evaluation methods performance is sub-par. The amount of variation in polysynthetic words indicates that choosing a limited amount of words solely based on tf-idf metrics is not an optimal method.
As was shown when translation quality was mea-

Lang	Method	Precision(%)	Recall(%)	F1-score
EN-IU 1 word	Null-Dic	73.0	97.9	83.6
EN-IU 2 word	Null-Dic	49.9	97.9	66.1
EN-IU 3 word	Null-Dic	32.6	97.9	48.9
EN-IU 1 word	Segmented	88.8	97.9	93.1
EN-IU 2 word	Segmented	75.3	97.9	85.1
EN-IU 3 word	Segmented	63.4	97.9	77.0

Table 5.4: Results for English-Inuktitut parallel texts and segmented parallel texts measured by 1, 2 and 3 occurrences of translated words in a sentence

Lang	Method	Precision(%)	Recall(%)	F1-score
Da-kl 1 word	Null-Dic	64.6	98.0	77.86
Da-kl 2 word	Null-Dic	38.5	98.0	55.28
Da-kl 3 word	Null-Dic	27.2	98.0	42.6
Da-kl 1 word	Segmented	74.0	98.0	84.32
Da-kl 2 word	Segmented	62.3	98.0	76.2
Da-kl 3 word	Segmented	51.7	98.0	67.7

Table 5.5: Results for Danish-Kalaallisut parallel texts and segmented parallel texts measured by 1, 2 and 3 occurrences of translated words in a sentence

sured on English-Swedish and English-Finnish, word pairs that occurred less than 4 times were unreliable to be a good indication of word similarity. This lead to the majority of co-occurring words being deleted for the polysynthetic languages, and the ones being used having far less co-occurrences than words in English-Swedish and English-Finnish. The results do however show that this issue can largely be alleviated by the use of morphological segmentation, where results come much closer to that of non-polysynthetic sentence alignment.

It was also thought that the alignment quality of the different polysynthetic languages would show similar results, however this was not the case. It is not entirely clear why the Danish-Kalaallisut scores were so much worse than that of English-Inuktitut. It could be due to more noise being present in the magazine chosen for Danish-Kalaallisut which was printed in 1999, although this is difficult to confirm. Another possible explanation is that the abundance of compound words found in Danish, in comparison to those found in English, resulted in words co-occurring even less frequently between Danish and Kalaallisut, as the change in morphology of words given the sentence structure in kalaallisut made any translations of compound words to infrequent to be considered reliable, even when performed on segmented data.

Overall it can be said that the method of evaluation is not robust enough. From looking at the sentences in all languages where zero co-occurring words were found, it seems that many of these sentences are properly aligned. Usually this occurred in shorter sentences such as article titles. The method might work better with a different way of identifying translations and not doing it fully automatically. Taking more care to include words from shorter sentences and titles would likely improve results.

In summary, as there was no gold standard or native speakers available to evaluate the systems performance for the polysynthetic languages, it was neccessary to find a more automatic method to evaluate alignment quality. It seems the method of evaluation introduced in this text has a certain level of merit to it, and that the use of co-occurrence metrics is a good choice in establishing context based translation pairs between languages, but that the use of tf-idf may not be the optimal method to choose what words should be included in the word lists. Better methods of word accumulation need to be looked into.

It is also shown that morphological segmentation can serve to greatly improve results, but it is uncertain if these results are indicative of improvement in actual quality of alignment or just in making the alignment quality easier to evaluate.

5.6 Bitextor and Machine Translation Evaluation

Part of this project was done in combination with the development of neural machine translation (NMT) systems submitted by the University of Groningen to the English–Inuktitut language pair WMT 2020 translation tasks. As such, the data gathered and used could be evaluated by these NMT systems. The Danish Kalaallisut data was also used to see if the data of a closely related language to Inuktitut would improve the outcome performance of the systems developed. Danish was however first translated to English with a pretrained DAEN system from OPUS-MT (Tiedemann and Thottingal, 2020) before being used by the system. It should be noted here that the data gathered from the use of Bitextor was significantly reduced at this stage to 14778 sentence pairs after removing any sentences pairs not classified as con-

taining Danish. Danish-Kalaallisut was also evaluated separately using the NMT systems on different morphological segmentation models developed by Mol (2020).

System	IU→ EN	EN→ IU
Best constrained (5, 3 resp.)	**22.24**	51.31
+ IU magazine	22.22	**51.88**
+ IU mag + KL mag		50.57
+ IU mag + KL crawl		51.27

Table 5.6: Results of the unconstrained systems for both translation directions and both dev sets. The scores are measured in BLEU (IU→EN) and CHRF (EN→IU). Best results are shown in bold. Taken from (Roest et-al, 2020)

MODELNAME	VALIDATION	TEST
no segmentation	- 2.14 BLEU	2.05
flatcat	- 1.67 BLEU	1.53
lmvr	- 1.34 BLEU	1.11
morfessor	- 1.27 BLEU	0.95
bpe 4k	- 1.26 BLEU	1.60
bpe 20k	- 1.18 BLEU	1.44
crf	- 0.94 BLEU	1.01

Table 5.7: Results of Kalaallisut → English translation as performed by NMT system based on the one submitted by the University of Groningen for the 2020 WMT task

The results for from the NMT task can be seen in table 5.6 The results for NMT translation from English to Kalaallisut can be seen in table 5.7
The results from the English translation are mixed. Table 5.6 shows that CHRF improved by 0.55 points when translation done from English to Inuktitut included aligned English-Inuktitut magazines in the training data. No improvement occurred however when translation was done from Inuktitut to English. Table 5.6 also shows that the use of English and Kalaallisut bitextor data and aligned magazines marginally worsened performance of the system by 0.04 points. This could be the result of to little data being used resulting in noise being added to the system. The NMT system translation scores for English to Kalaallisut found in table 5.7 are very poor. As the system only used post-processed Bitextor data and magazine articles it is likely that the systems performance simply did not have enough data.
The data used from bitextor was also very noisy and did not always produce very accurate translations, as indicated by almost 90% of the sentences having to be deleted before usage by the NMT system.
The use of Bitextor was very limited due to a lot of parameters not being able to be taken advantage of, as they required lexical or otherwise external information about the language. This type of information was not available for Kalaallisut at the time of doing this project.

6 Conclusion

This project attempted to find out if good quality sentence alignment can be performed on low-resource polysynthetic languages and if a method of evaluation using tf-idf and co-occurrence statistics can be incorporated to determine quality of alignment when no gold standard is available.
While alignment techniques do seem to be applicable to polysynthetic languages and can be helpful to improve translation quality in NMT system (see table 5.6), a lot of problems are raised in the attempt to evaluate the alignment quality. The proposed method of evaluation was shown not to accurately reflect actual quality of alignment.
The project also sought to find out if methods of automatic parallel text retrieval could be applied to provide large amounts of parallel data to be used by existing NMT systems (see tables 5.6). While initially seeming very promising, the complexity of the language and lack of available external resources for Kalaallisut, resulted in very noisy data and ended up not being helpful in the improvement of existing NMT systems.

References

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, and Ludovic Rheault. Digitization of the canadian parliamentary debates. *Canadian Journal of Political Sci-*

ence/Revue canadienne de science politique, 50 (3):849–864, 2017.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

Miquel Espla-Gomis and Mikel Forcada. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93(2010):77–86, 2010.

William A Gale and Kenneth Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

Joseph H Greenberg. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3): 178–194, 1960.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *LREC*, pages 489–492, 2006.

Chenggang Mi, Yating Yang, Xi Zhou, Xiao Li, and Turghun Osman. Co-occurrence degree based word alignment: A case study on uyghur-chinese. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 259–268. Springer, 2014.

Marianne Mithun. Polysynthesis in the arctic. *Variations on polysynthesis: The Eskaleut languages*, pages 3–17, 2009.

Barbera de Mol. A comparison of data-driven morphological segmenters for low-resource polysynthetic languages: A case study of greenlandic (unpublished bachelor thesis). 2020.

Xiaojun Quan, Chunyu Kit, and Yan Song. Non-monotonic sentence alignment via semisupervised learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 622–630, 2013.

C. Roest. Machine translation for english–inuktitut with segmentation,data acquisition and pre-training (unpublished master thesis). 2020.

Danica Seničić and Pr Cédrick Fairon. Automatic alignment of bilingual sentences. 2017.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*, 2006.

Jörg Tiedemann. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165, 2011.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012.

Antonio Toral, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way. Crawl and crowd to bring machine translation to under-resourced languages. *Language resources and evaluation*, 51 (4):1019–1051, 2017.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247, 2007.

Krzysztof Wołk and Krzysztof Marasek. A sentence meaning based alignment method for parallel text corpora preparation. *Advances in Intelligent Systems and Computing, Springer*, 275: 229–239, 04 2014. doi: 10.1007/978-3-319-05951-$8_2$2.

Z. Zhandos, M. Aigerim, and R. Diana. New kazakh parallel text corpora with on-line access. In *International Conference on Computational Collective Intelligence*, pages 501–508. Springer, 2017.