Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2562–2572
Marseille, 11–16 May 2020

The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine Translation Results

Eric Joanis, Rebecca Knowles, Roland Kuhn
Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart
National Research Council Canada
1200 Montreal Road, Ottawa ON, K1A 0R6
firstname.lastname@nrc-cnrc.gc.ca, chikiu.lo@nrc-cnrc.gc.ca

Jeffrey Micher
US Army Research Laboratory
2800 Powder Mill Road, Adelphi MD 20783
jeffrey.c.micher.civ@mail.mil

Abstract

The Inuktitut language, a member of the Inuit-Yupik-Unangan language family, is spoken across Arctic Canada and noted for its morphological complexity. It is an official language of two territories, Nunavut and the Northwest Territories, and has recognition in additional regions. This paper describes a newly released sentence-aligned Inuktitut–English corpus based on the proceedings of the Legislative Assembly of Nunavut, covering sessions from April 1999 to June 2017. With approximately 1.3 million aligned sentence pairs, this is, to our knowledge, the largest parallel corpus of a polysynthetic language or an Indigenous language of the Americas released to date. The paper describes the alignment methodology used, the evaluation of the alignments, and preliminary experiments on statistical and neural machine translation (SMT and NMT) between Inuktitut and English, in both directions.

Keywords: Inuktitut, SMT, NMT, sentence alignment, machine translation for polysynthetic languages, Indigenous languages

1. Introduction

This paper describes the release of version 3.0 of the Nunavut Hansard Inuktitut–English Parallel Corpus, a sentence-aligned collection of the proceedings of the Legislative Assembly of Nunavut (Nunavut Maligaliurvia).[1] The Nunavut Hansard 3.0 corpus consists of 8,068,977 Inuktitut words and 17,330,271 English words and covers the proceedings of the 687 debate days from April 1, 1999 to June 8, 2017.

We include an account of several automatic sentence alignment techniques examined, along with a summary of experiments in which gold standard sentence alignments created by Inuktitut–English bilinguals were compared with those created by our chosen automatic alignment technique. We also present baseline statistical and neural machine translation experiments on the corpus, showing the positive effects of improved alignment and corpus size increase, and provide a brief discussion of challenges of machine translation into morphologically complex languages. This corpus release marks a move away from machine translation "low-resource" status for this language pair and domain.

2. Background

2.1. Nunavut and Inuktut

Nunavut is a territory of Canada that comprises most of the Canadian Arctic archipelago. Separated from the Northwest Territories on April 1, 1999, it is the largest of Canada's territories. According to the 2016 census, 65.3% of the population of Nunavut (of 35,944) had a variety of Inuktut (Inuktitut or Inuinnaqtun) as their first language, and 89.0% of Inuit in Nunavut could speak Inuktut at a conversational level or above (Lepage et al., 2019).

Figure 1: The Nunavut territory and its location within Canada. Map data ©2019 Google, INEGI

Inuktut (not to be confused with *Inuktitut*, a more specific term) refers to a group of dialects spoken in Nunavut that includes Inuktitut and Inuinnaqtun; it and other Inuit languages form a dialect continuum that stretches from Alaska to Greenland. The transcription conventions used in this corpus tend towards North Qikiqtaaluk (Baffin Island) forms, but as speakers in the Legislative Assembly (as well as their interpreters and transcribers) have diverse linguistic backgrounds, users of this corpus should not necessarily take any given word or sentence to be representative of any particular variety of Inuktitut.

Inuktut is known for having considerable morphological complexity. Words are typically multi-morphemic, with roots followed by multiple suffixes/enclitics from an inventory of hundreds of possible elements. Moreover, these el-

[1] The corpus is available at the NRC Digital Repository: https://doi.org/10.4224/40001819 (CC-BY-4.0 license).

ements undergo complex (albeit regular) morphophonemic changes in contact with each other, obscuring the underlying form of many morphemes (Micher, 2017). Here is a typical example sentence from this corpus:

ᐅᖃᖅᑎ, ᑌᓐᓇᒃᑯᓪ ᐊᐅᓪᓴᐃᑎᐊᓪᓚᑦᑕᕋᓂᒪᒐ ᐅᖃᐅᓂᖅ ᐅᒪᓇᖅᑎᑕᐅᖅᑲᓂᕋᓂᓪᓗᓂ ᐊᒪᒪᓗ ᐃᓕᓴᕐᓂᖅᑎᑦ ᐃᓕᓴᕐᓂᖅᑎᑕᐅᓗᑎᒃ.

uqaqtii, gavamakkut aulajjaigiallattaarasummata uqausiq uummaqtitaukkannirasulluni ammalu ilinniaqtiit ilinniaqtitaulutik.

Mr. Speaker, this government is taking concrete action on language revitalization and student outcomes.

This complexity contributes to the scientific interest of this corpus, raising questions like whether contemporary subword methods in natural language processing (Sennrich et al., 2015; Kudo and Richardson, 2018, among others) are sufficient to process one of the world's most morphologically complex languages, and if not, how to integrate prior knowledge of Inuktitut morphology in a robust way. The work described in this paper can also be viewed in the context of a wide range of language technologies currently being developed for Indigenous languages spoken in Canada, as surveyed by Littell et al. (2018).

2.2. Legislative Assembly of Nunavut and Hansard

The Legislative Assembly of Nunavut conducts business in both Inuktut and English, with the choice of language influenced by both Member preference and the topic at hand (Okalik, 2011), and its proceedings, the Nunavut Hansard, are published in both Inuktitut and English. "The policy of Nunavut's Hansard editors is to provide a verbatim transcript with a minimum of editing and without any alteration of the meaning of a Member's speech."[2]

Prior to approximately 2005–2006,[3] the English version of the Hansard was a transcription of the English speeches when English was spoken, and a transcription of the live English interpretation when Inuktitut or Inuinnaqtun was spoken. The Inuktitut version was a later translation of the English version; the Inuktitut sentences in these versions should correspond in meaning to what was said but will not exactly match the sentences as spoken in the Assembly.

2005–2006 marked a change in the production of the Hansard; now both the English and Inuktitut versions are transcriptions of the speeches or their live interpretations into the respective languages (Riel Gallant, personal communication, Sept. 6, 2019). Speeches in other varieties of Inuktut are also interpreted into Inuktitut when they are not mutually intelligible, and it is these live interpretations that are transcribed in the Hansard. The Hansard contains parenthetical annotations of when interpretation occurs (e.g., "interpretation begins", "interpretation ends"). When the transcriber considers the interpretation to contain an error (e.g., missing information), they will instead translate the original sentence into the target language.

[2] `assembly.nu.ca/hansard`

[3] We are unable to confirm the exact date at which this change in production occurred, or whether it happened gradually.

2.3. Prior Releases and Work on Alignment

The first Nunavut Hansard parallel corpus was released to the research community in 2003 (Martin et al., 2003). Nunavut Hansard 1.0 (NH 1.0), covered 155 days of proceedings of the Nunavut Assembly, from April 1, 1999 to November 1, 2002; it comprised 3,432,212 English tokens and 1,586,423 Inuktitut tokens. Sentence alignment was carried out by a modified version of the Gale and Church (1993) algorithm. An improved release, NH 1.1, was one of three corpora used as experimental data for a 2005 shared task on word alignment for languages with scarce resources (Martin et al., 2005).

Version 2.0 of the corpus (NH 2.0) was released in January 2008. It covered the proceedings of the Nunavut Assembly through November 8, 2007 (excluding 2003). NH 2.0 contained 5,589,323 English words and and 2,651,414 Inuktitut words. Sentence alignment was carried out with Moore's aligner (Moore, 2002).[4]

For this work, we use our in-house sentence aligner, based on Moore (2002) with some enhancements. Like Braune and Fraser (2010) and most sentence aligners, we allow 1–many alignments, and like Yu et al. (2012) we use a multi-pass approach to refine alignments.

A recent approach to sentence alignment (Thompson and Koehn, 2019) is reported to significantly improve alignment quality by incorporating bilingual word embeddings trained on large pre-existing parallel corpora. We experimented with this approach as well (see §3.3.).

3. Alignment of the Nunavut Hansard Inuktitut–English Parallel Corpus 3.0

The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (NH 3.0) consists of 17,330,271 English words in 1,452,347 sentences, and 8,068,977 Inuktitut words in 1,450,094 sentences, yielding approximately 1.3 million aligned sentence pairs.[5] It covers the proceedings through June 8, 2017.

3.1. Text Extraction from Word Documents

We received the corpus as a collection of Word documents graciously provided by the Legislative Assembly of Nunavut, from which we then extracted the text.

The Inuktitut transcriptions of sessions 1–5 of the 1st Assembly are encoded in the ProSyl/Tunngavik font over Roman characters. While such fonts enabled the use of syllabics (as well as IPA and other symbols) prior to their inclusion in character sets like Unicode, they make text extraction more difficult today. Applying standard text extraction tools to these fonts will produce jumbled sequences of letters, digits, and punctuation rather than syllabic characters. For example, "Legislative Assembly of Nunavut" in Inuktitut would be extracted as "kNK5 moZos3=x" instead of "ᓄᓇᕗᑦ ᒪᓕᒐᓕᕆᔨᒃ" (or, in transliteration, "nunavut

[4] NH 1.0/1.1 and NH 2.0 were preprocessed and sentence-aligned by a team at the National Research Council of Canada (NRC). Subsequently, a member of that team, Benoît Farley, set up a site from which these corpora and other resources can be downloaded: `www.inuktitutcomputing.ca`.

[5] Sentences may get merged during the alignment process when translation is not done in linear, one-to-one correspondence.

maligaliurvinga"). For this corpus, which contains some contents in English on the Inuktitut side (written in standard fonts, which should therefore not be converted), conversion could not be done as a simple post-extraction step. Lossy heuristics or language/character identification would be required to recover information that was explicitly encoded in the Word documents via the font information.

For those Inuktitut documents, the text was extracted using the IUTools toolkit from Inuktitut Computing, which iterates over all the text strings in the Word document with an API that provides the font information for each string. When the font for a given piece of text was labelled as ProSyl, the text was converted according to the Tunngavik font tables,[6] with some heuristics to correct errors.[7] The output is plain text in UTF-8 with one paragraph per line.

For the other sessions and for all English documents, we used AlignFactory[8] (AF) to extract the collection of paragraphs out of the Word documents.[9] There are many open-source tools that can do the same task with similar results, but we find this commercial tool reliable and easy to use. AF creates a TMX file from which we extract the text to plain text files in UTF-8 with one paragraph per line.

Segmentation of the paragraphs in both languages into sentences was done using the Portage (Larkin et al., 2010) sentence splitter, which is part of the Portage tokenizer.

To remove parts of the corpus that were clearly not parallel, we searched for outliers using various heuristics: documents with unusually large or small paragraph or text length ratios between the languages; with similar imbalances in the appendices; or where a first pass of the baseline alignment methodology (see §3.3.) had unusually low alignment probability. These outliers were visually inspected and non-parallel text was manually removed. Examples of problems found include: an appendix occurring only in one language; an appendix occurring as text in one language but as scanned images from which we cannot extract text in the other; the list of Members of the Legislative Assembly (MLAs) being in a different order between the two languages or in a different place in the documents. We also performed automated cleanup to remove error and logging messages from the text extraction software. Finally, we standardized some UTF-8 space, hyphen, and punctuation characters to produce cleaner plain text files.

The syllabic characters are not normalized in any way; they are left as found in the original documents or as converted from the ProSyl/Tunngavik font. We note that the data contains some syllabic characters from other Canadian Indigenous languages, which are not part of the Inuktitut alphabet; we do not attempt to modify or remap these. We do not perform any orthographic normalization or spell checking (on either the English or Inuktitut side of the corpus). The intent is to maintain the parallel parts of the corpus as close to its original published version as possible. In §4.2. we discuss some potential ways of handling unexpected characters in the text for machine translation experiments.

Date	Paragraphs EN	Paragraphs IU	Sentences EN	Sentences IU
1999-04-01	489	420	747	706
2001-02-21	528	540	1,134	1,144
2002-10-29	566	572	1,074	1,078
2006-11-23	1,019	1,030	2,259	2,161
2013-03-13	1,640	1,741	3,729	3,399
2016-10-24	1,480	1,524	3,148	3,279

Table 1: Size of each gold standard debate day. Across the whole corpus, each debate day averaged about a thousand paragraphs and two thousand sentences in each language.

3.2. Gold Standard and Evaluation Methodology

The goal of this work is to produce a well-aligned corpus. In order to measure how well we are able to align the corpus, we compare our various alignment methods (described in §3.3.) using gold-standard alignments on a subset of the corpus (six debate days).

Martin et al. (2003) used three debate days to evaluate NH 1.0. While we were unable to obtain the original gold standard they used, we selected their three evaluation debate days (1999-04-01, 2001-02-21, and 2002-10-29) as half of the six we used to create a new gold standard. To these, we added three days spread roughly evenly over the new parts of the corpus, one per Assembly since the first: 2006-11-23, 2013-03-13, and 2016-10-24. This gave us a total of approximately 12,000 sentences of text to manually align. Table 1 shows the sizes of each of the debate days contained in the gold-standard alignment set.

To produce the gold-standard alignments, we hired two Inuktitut language experts (both first-language speakers) through the Pirurvik Centre.[10] The annotators were asked to work in InterText Editor[11] to produce the sentence alignments from pre-segmented text: we provided the sequence of sentences and they produced the alignments. To avoid biasing the annotators, we did not run any automated pre-alignment (as InterText normally recommends doing).

To evaluate the reliability of the gold standard, we follow Li et al. (2010) and compute the inter-annotator agreement (IAA) on the alignments using F-score over mutual precision and recall, which is mathematically equivalent to $\frac{2*|\mathcal{G}_1 \cap \mathcal{G}_2|}{|\mathcal{G}_1|+|\mathcal{G}_2|}$, considering one annotator's set of alignment links ($\mathcal{G}_1$) as the reference and the other ($\mathcal{G}_2$) as the measured, or vice versa. The overall IAA is 93.1%. The first row in table 2 shows IAA by debate day and overall. Some

[6] Syllabic characters in the Word documents were marked as being in the ProSyl font, but were in fact in the Tunngavik font, an extension of ProSyl.

[7] For example, the dot representing long vowels is sometimes present multiple times in the text (which is invisible in print: the ProSyl/Tunngavik dots were non-advancing characters, so multiple super-imposed dots appeared as one dot), and it is not always in the right position. In Unicode, there is only one possible dot for each syllabic character, so any sequence of dots associated with a character was normalized to the correct dot for that character.

[8] `terminotix.com/index.asp?content=item&item=4`

[9] Although AF is also designed for alignment, we do not use it in our experiments because its alignment algorithm relies on assumptions that do not generalize well to Inuktitut–English text alignment. However, it extracts the text well.

[10] `www.pirurvik.ca`

[11] `wanthalf.saga.cz/intertext`

Annotation set	1999-04-01	2001-02-21	2002-10-29	2006-11-23	2013-03-13	2016-10-24	Overall
Independent	93.2	95.2	98.0	89.7	92.5	93.9	93.1
Consensus-corrected	96.4	97.9	100.0	97.1	97.5	98.5	97.9

Table 2: Inter-annotator agreement (%) on the gold standard sentence alignments by debate day and overall, calculated on the original independent annotations and on the consensus-corrected ones

debate days are clearly more difficult to align than others, and that is also reflected in the alignment results (§3.4.).

In order to produce a high quality gold standard, we asked the annotators to review differences between their alignments, categorizing those differences as one of the following types: both reasonable; one correct; the other correct; both need fixing. There were 483 blocks of alignments in the consensus surveys, involving a total of 1160 Inuktitut sentences and 1343 English ones. 18% of the blocks were marked as both reasonable, 74% as having only one correct annotation, and 8% as both incorrect. We also provided a free text box for additional comments and clarifications. Using this box, 41% of survey blocks were flagged as having incomplete or incorrect translations. This reflects the fact that the corpus is not strictly parallel; some text is not fully translated, some is reordered in translation, and verbose prose is sometimes summarized.

We created the consensus-corrected gold-standard annotations by correcting the independent annotations only in those cases where the consensus answer was that only one annotation was correct, leaving both-acceptable and both-need-fixing cases unchanged. The second row in Table 2 shows consensus-corrected IAA, which is not 100% because of the cases where consensus did not yield a single correct alignment.

The consensus-corrected gold standard annotations are used in all evaluation results reported below. Both the independent annotations and the consensus-corrected ones are distributed with the corpus, to support future work on improving the alignment of this corpus.

To evaluate the quality of the automated alignment of the corpus, we follow Och and Ney (2003) and calculate the alignment error rate (AER) of the predicted alignment ($\mathcal{A}$). AER is a measure based on the F-score, calculating recall over Sure links ($\mathcal{G}_S \equiv \{\mathcal{G}_1 \cap \mathcal{G}_2\}$, i.e., the set of agreed links among the two consensus-corrected annotation sets) and precision over Possible links ($\mathcal{G}_P \equiv \{\mathcal{G}_1 \cup \mathcal{G}_2\}$, i.e., the set of all collected links from the two consensus-corrected annotation sets): $\text{AER} = 1 - \frac{|\mathcal{A} \cap \mathcal{G}_S| + |\mathcal{A} \cap \mathcal{G}_P|}{|\mathcal{A}| + |\mathcal{G}_S|}$.

3.3. Alignment Experiments

We perform alignment experiments comparing several alignment approaches and different whole-word, subword, and morphological vocabularies. The results are discussed in §3.4., and presented in Table 3.

Our baseline approach is a 4-pass alignment methodology, using `ssal`, a reimplementation of Moore (2002) used in the Portage machine translation system (Larkin et al., 2010). Unlike Moore (2002), we use a 4-pass rather than 2-pass approach, as follows:

1. align paragraphs using dynamic programming over paragraph lengths, based on Gale and Church (1993),
2. align sentences within each aligned paragraph pair using sentence lengths,
3. train an IBM-HMM model (Och and Ney, 2003) on the output of 2. and use it to re-align paragraphs, as Moore (2002) does,
4. align sentences within aligned paragraphs using the IBM model again.

This 4-pass approach is an extension of the 3-pass approach implemented in `ssal` by Foster (2010). `ssal` optimizes the dynamic programming search by starting in a narrow diagonal and widening the search only as necessary, and allows 0–1, 1–1, 1–many and many–many alignments, with a maximum of 3 units on either side of an aligned pair.

In past work on sentence alignment, we have found that first aligning paragraphs and *then* aligning sentences within paragraphs outperforms approaches that align sentences without paying attention to paragraph boundaries. To demonstrate the strength of this 4-pass approach, we also present results on 1-pass and 2-pass approaches. We report on the following experiments (all done using `ssal`):

- **Gale+Church 1 pass:** only length-based alignment over sequences of sentences with no concept of paragraph boundaries.
- **Gale+Church 2 pass:** steps 1 and 2 of the 4-pass methodology described above.
- **Moore-style 2 pass:** length-based and then IBM-HMM models, over sequences of sentences with no concept of paragraphs.

In our 4-pass baseline and the 1- and 2-pass ablation experiments, the corpus is only tokenized on punctuation, using the Portage tokenizer, but words are kept whole. For English, this generally works well. For polysynthetic languages like Inuktitut, with high type-token ratios, data sparsity presents a major challenge for statistical models. In order to experiment with ways of overcoming data sparsity for alignment, we perform experiments on subword and morphological segmentations of the text.

The morphological experiments (**Morpho surface** and **Morpho deep**) use morphological segmentation, replacing Inuktitut words with either sequences of surface forms (simple morphological word segmentation) or sequences of deep forms (segmentation with substitution). The morphological analysis was done on romanized words using Uqailaut (Farley, 2009). Words that had no analysis from the Uqailaut analyzer were subsequently processed with the neural analyzer described by Micher (2017) and Micher (2018). While the Uqailaut analyzer produced multiple analyses per word, only the first analysis was used in these experiments. Furthermore, as the neural analyzer only produces a 1-best

analysis for each word, each word in the full corpus has a single, 1-best analysis.[12] We follow the same 4-pass alignment methodology as in the baseline, but train and apply the IBM-HMM model on sequences of surface or deep forms instead of words.

The morphological segmentation experiments rely on the existence of a morphological analyzer, a tool that is not available for all languages. To demonstrate that simple heuristics can also perform well in the absence of morphological analysis tools, we performed experiments using the **Prefix** of each word as a stand-in for more complicated lemmatization (Simard et al., 1992; Och, 2005; Nicholson et al., 2012). As a polysynthetic language, Inuktitut has many very-low frequency words, and even English words have some morphology which a stemmer could normalize. Since syllabic characters represent whole syllables, we chose a prefix of three characters to approximate the stem of an Inuktitut word and we use a five letter prefix for English words. We follow the same alignment methodology as in the baseline, but train and apply the IBM-HMM model on sequences of prefixes instead of words.

Byte-pair encoding (BPE) is another approach to word segmentation that does not rely on existing language-specific tools (Sennrich et al., 2015). BPE is a compression algorithm often used in neural machine translation to build fixed-size vocabularies that nevertheless allow for all words (within a given character set) to be represented as sequences of subword units from the vocabulary. This simple method has the disadvantage that it can break words into semantically meaningless substrings, but it has the advantage that it only keeps the most frequent sequences of characters, which plays well with statistical models like the IBM ones. We used `subword-nmt` (Sennrich et al., 2015) to train and apply the BPE models. We tried several vocabulary sizes, as well as joint and disjoint training. In the results below, we report only our first experiment, **BPE baseline (30k joint)**, and the one with the best results **BPE best (10k joint)**. We tried additional settings that we do not report: all have scores between the two reported here.[13]

We can also combine the linguistically informed morphological analysis with automatic methods like BPE, as we do in our **Morpho deep + BPE** experiment. The morphological segmentation with substitution of deep forms has the advantage of grouping together different surface forms of the same underlying morphemes for the statistical model, while BPE has the advantage of treating high frequency sequences of characters as units while breaking up lower frequency ones. We combine the two as follows: run morphological segmentation, replace the surface form by the deep form for each morpheme found, and glue the sequence of deep forms back together into pseudo-words; train and run BPE (10k joint) on those pseudo-words; use the results as in the other experiments in the IBM-HMM models.

Shortly before submission, **Vecalign** (Thompson and Koehn, 2019) reported a new state of the art in alignment, using on high-quality pre-trained bilingual word vectors ("bivectors"). Since pretrained bivectors for Inuktitut–English do not exist, to the best of our knowledge, we use `bivec` (Luong et al., 2015) to train bivectors on our best system output (BPE 10k joint), and then use the results to realign the whole corpus with vecalign.[14] The embeddings, of dimension 1024, are trained for 5 iterations using skip-gram, a maximum skip length between words of 11, 10 negative samples, a hierarchical softmax, and without discarding any words based on frequency counts. We create Vecalign overlaps using `overlaps.py` and an overlap window of 15 for the source and the target separately. Using `bivec`'s word embeddings, we average the word embeddings for each sentence in the overlap sets to produce their sentence embedding. We then align a pair of files with `Vecalign` using a maximum alignment size of 15, the overlaps, and the sentence embeddings of the overlaps.

3.4. Evaluation Results

Table 3 shows the results over our six gold standard sets for each of the methodologies, as well as overall. The overall AER is not the average of the six, but AER calculated on the concatenation of the six documents.

It should be noted that the alignment difficulty of each day in the gold standard varies considerably, in a way that inversely correlates with the IAA on each day, due largely to content that is not fully matching in both languages.

It is clear from the results that morphological (or pseudo-morphological) segmentation gives considerable improvements over the whole-word baseline; this is unsurprising given the morphological complexity of Inuktitut.

Automatic methods—even the relatively naïve method of reducing words to their first N characters—performed comparably with more principled methods of morphological analysis; overall, a simple BPE segmentation with a small vocabulary (10k joint) performed best, though only slightly better than other methods. We still consider true morphological methods to be promising avenues of future work: there are still many improvements that can be made in Inuktitut morphological analysis, whereas what we see here might be the ceiling of BPE-based approaches.

The Vecalign method performs relatively poorly, likely due to the absence of a large, independent corpus with which to pre-train Inuktitut–English word vectors. Like morphological methods, we still believe this to be a promising approach, as there are many ways to train bilingual word vectors (Ruder et al., 2019) that could lead to improvements.

3.5. Comparison with NH 1.0

In order to quantify the improvement of the alignments that NH 3.0 brings over the original NH 1.0 corpus, we compared AER scores on the three gold standard files that were included in both versions of the corpus. We do not have access to the original gold standard from NH 1.0, but we can

[12] Analyzed forms are available at `www.cs.cmu.edu/afs/cs/project/llab/www/`

[13] In an ideal world, we would have wanted a dev set to tune the BPE meta-parameters, but the desired output of this work is the best corpus possible and we only have a small amount of manually annotated data, so we simply retain the BPE settings that yield the best results over our gold standard.

[14] We tried several other settings for Vecalign, but report only the one that performed best.

Method	1999-04-01	2001-02-21	2002-10-29	2006-11-23	2013-03-13	2016-10-24	Overall
Gale+Church 1 pass	26.0	10.4	2.8	25.4	19.7	22.4	19.6
Gale+Church 2 pass	28.8	7.0	3.9	14.4	13.6	11.8	12.7
Moore-style 2 pass	18.1	6.4	1.4	18.1	13.7	15.5	13.6
Baseline (4 pass)	20.7	5.1	2.8	11.8	10.7	10.4	10.2
Morpho surface	16.4	3.4	1.4	9.2	8.1	8.2	7.8
Morpho deep	**11.7**	**2.9**	1.3	8.1	7.4	7.6	6.9
Prefix	12.6	3.0	1.4	8.8	7.5	7.9	7.3
BPE baseline (30k joint)	16.2	3.5	1.2	8.9	7.6	8.2	7.6
BPE best (10k joint)	13.8	3.5	1.2	**7.3**	**6.7**	**7.2**	**6.6**
Morpho deep + BPE	13.5	3.7	**1.0**	7.5	7.4	7.4	6.9
vecalign	14.5	6.7	5.1	11.4	11.1	11.4	10.5

Table 3: Alignment error rates (%) for each gold standard debate day, and overall, for each alignment methodology

Method	1999-04-01	2001-02-21	2002-10-29	Overall
NH 1.0	16.0	11.5	7.2	11.0
NH 3.0 baseline	17.5	5.2	2.9	7.3
NH 3.0 best (BPE 10k)	**10.0**	**3.4**	**1.3**	**4.2**

Table 4: Alignment error rates (%) on the subset of sentences in both NH 1.0 and NH 3.0

use our new gold standard. Unfortunately, the text extraction was not done in the same way for both versions.[15]

In order to make NH 1.0 and NH 3.0 comparable despite the differences in text extraction, we extracted the subset of sentences that occur in both NH 1.0 and NH 3.0 and calculated the AERs for the three relevant gold standard files over that subset of each document.

We see in Table 4 that our new version of the corpus not only adds a large amount of new text to the corpus, but also significantly improves the quality of the alignments, with the AER reduced from 11.0% to 4.2%.

4. Machine Translation

To illustrate the value of this corpus in machine translation, and the quality improvements that the much larger corpus and improved alignment make possible, we offer baseline results using statistical (SMT) and neural (NMT) machine translation systems. We describe the data splits (§4.1.) and preprocessing (§4.2.), our SMT and NMT setups (§4.3.), and baseline MT results (§4.4.) with a brief discussion of some challenges of English–Inuktitut MT and evaluation.

4.1. Training, Development, and Evaluation Splits

For development and evaluation, we release a development set (dev) a development-test set (devtest) and a held-out test set (test) constituted of the most recent material in the corpus, keeping everything else in the training set (train). These sets respect document boundaries, with dev containing 2017-03-14, 2017-05-30 and 2017-05-31, devtest containing 2017-06-01, 2017-06-02 and 2017-06-05, and test containing 2017-06-06, 2017-06-07 and 2017-06-08. We release both a full version of these sets, which could be used for the evaluation of document-level machine translation, and a deduplicated version. As a parliamentary corpus, the Hansard contains extremely high levels of repetition, which could reward simple memorization of the training corpus. In order to avoid this, our machine translation experiments described below report scores on the deduplicated version of the evaluation splits. The deduplication was performed as follows: for each debate day in dev, devtest, and test, any pair of English–Inuktitut sentences that had appeared (exact match) at any earlier date in the corpus was removed. The deduplicated debate days contain on average 44% fewer sentence pairs than the full debate days.

4.2. Preprocessing

We used consistent preprocessing across all of our machine translation experiments, to allow for as accurate comparisons as possible. We release these preprocessing scripts with the corpus for the purpose of replicability.

We first convert Inuktitut data in syllabics to romanized form (as prior MT work like Micher (2018) has done) using `uniconv`.[16] To repair output errors (e.g., of accented French language characters appearing in the Inuktitut data) we then passed the Inuktitut side of the data through `iconv`. There remained several characters that were not handled correctly (certain punctuation and additional syllabics); these were identified and then corrected with an additional preprocessing script.[17] Word-internal apostrophes were treated as non-breaking characters on the Inuktitut side of the data (and thus not tokenized).

[15] Cursory visual inspection suggests that the text is better extracted in NH 3.0, but we did not attempt to quantify that. One obvious example is that in NH 1.0, some paragraph boundaries, e.g., in some bullet points or table cells, were lost without even a space, resulting in the concatenation of separate actual words into apparently very long Inuktitut words.

[16] `uniconv` is distributed with Yudit: `www.yudit.org`

[17] Note that this romanization pipeline does not fully conform to all spelling and romanization conventions described in the Nunavut Utilities plugins for Microsoft Word (`www.gov.nu.ca/culture-and-heritage/information/computer-tools`); we use the pipeline described here solely for MT experiments.

Following conversion to romanized script, we ran identical preprocessing with English defaults on both the Inuktitut and English sides of the corpus using Moses (Koehn et al., 2007) scripts: punctuation normalization, tokenization (with aggressive hyphen splitting), cleaned the training corpus (sentence length ratio 15, minimum sentence length 1, maximum 200), trained a truecaser on the training data and then applied it to all data. We trained byte-pair encoding models on the training data using `subword-nmt`,[18] with disjoint vocabularies ranging from 500 to 60,000 operations, which were then applied to all data. Future work may wish to consider promising morphologically informed approaches to segmentation for machine translation or alignment (Ataman et al., 2017; Ataman and Federico, 2018; Micher, 2018; Toral et al., 2019).

4.3. Baseline Experiments

We present baseline NMT and SMT results, over several versions of the data, in the hopes of encouraging future work on English–Inuktitut machine translation using this corpus. BLEU scores (Papineni et al., 2002) are computed with lowercase, v13a tokenization, using `sacrebleu` (Post, 2018). It should be emphasized that these scores are as high as they are in large part due to the formulaic and self-similar nature of the parliamentary genre, and should not be taken as representative of general-domain MT performance.

4.3.1. Neural Machine Translation

Our baseline neural machine translation system uses the Transformer architecture (Vaswani et al., 2017) implemented in Sockeye and following similar parameter settings to those described in Hieber et al. (2017): a 6-layer encoder, 6-layer decoder, a model dimension of 512 and 2048 hidden units in the feed-forward networks. As our byte-pair vocabularies were disjoint, we did not use weight-tying.

The network was optimized using Adam (Kingma and Ba, 2015), with an initial learning rate of 10^{-4}, decreasing by a factor of 0.7 each time the development set BLEU did not improve for 8,000 updates, and stopping early when BLEU did not improve for 32,000 updates.

We experimented with BPE vocabularies with 0.5, 1, 2, 5, 10, 15, 20, 25, 30, and 60 thousand merges. The maximum sentence length was set to 200, allowing us to compare the vocabularies without filtering out large numbers of training lines for the smaller vocabulary experiments. On the basis of BLEU and chrF (Popović, 2015) results on dev and devtest, we selected 2,000 merges for all reported English–Inuktitut NMT experiments, and 5,000 merges for Inuktitut–English NMT experiments. Translating into English, 5,000 merges produced the best results (or tied for best) on both metrics. Into Inuktitut, 2,000 merges tied for the best result in terms of BLEU and chrF on dev and had the best BLEU and second-best chrF on devtest, with very close performance to systems ranging from 500 to 10,000 merges. These results by BPE merges are in line with the low-resource Transformer results described by Ding et al. (2019), though NH 3.0 falls between their low-resource and high-resource settings.

[18] `github.com/rsennrich/subword-nmt`

4.3.2. Statistical Machine Translation

We trained our statistical machine translation system using Portage (Larkin et al., 2010), a conventional log-linear phrase-based SMT system. The translation model uses IBM4 word alignments (Brown et al., 1993) with grow-diag-final-and phrase extraction heuristics (Koehn et al., 2003). A 5-gram language model was trained on the target-side of the corpus using SRILM (Stolcke, 2002) with Kneser–Ney interpolated smoothing, modified when possible. The SMT system also includes a hierarchical distortion model with a maximum phrase length of 7 words, a sparse feature model consisting of the standard sparse features proposed by Hopkins and May (2011) and sparse hierarchical distortion model features proposed by Cherry (2013), and a neural network joint model (NNJM), with 3 words of target context and 11 words of source context, effectively a 15-gram language model (Vaswani et al., 2013; Devlin et al., 2014). The parameters of the log-linear model were tuned by optimizing BLEU on the development set using the batch variant of the margin infused relaxed algorithm (MIRA) by Cherry and Foster (2012). Decoding uses the cube-pruning algorithm of Huang and Chiang (2007) with a 7-word distortion limit. We report the average results from five tuning runs with different random seeds.

We ran SMT experiments with the same BPE vocabularies as used in NMT. On the basis of BLEU and chrF results on dev and devtest for NH 3.0, we selected 30,000 merges for all reported English–Inuktitut SMT experiments, and 5,000 merges for Inuktitut–English SMT experiments.

4.4. Results and Evaluation

We experiment with variations on the corpus to understand the effects of data size, recency, and alignment quality on machine translation output. In keeping with prior work, most of our results are on romanized (transliterated) Inuktitut, but we also present results on Inuktitut syllabics (**NH 3.0 syllabics)**. In all cases, as shown in Table 5, we report results on the same dev/devtest/test splits. The **NH 3.0** corpus is the full corpus, as aligned with the best alignment (BPE best 10k). Our best translation results, in both directions, occur with this dataset.

To measure the impact of alignment quality, we also examine **NH 3.0 beta**, the version of the corpus aligned with the baseline 4-pass alignment. We find that improved alignment generally improves translation quality as evaluated by automatic metrics. Switching from the Baseline (4-pass) alignment to the BPE best (10k joint) alignment, with an overall alignment error rate drop from 10.2% to 6.6%, results in BLEU score improvements between 0.3 and 0.4 (into English) and between 0 and 0.3 (into Inuktitut) for NMT, with no clear improvement for SMT, which is less sensitive to alignment noise than NMT (Khayrallah and Koehn, 2018).

While we would have liked to compare directly to the 1.0/1.1 and 2.0 versions of the corpus, they were romanized differently than our current corpus, so fair comparison is not possible. In order to examine the effect of the growth of the corpus, we selected subsets of the NH 3.0 training corpus, starting from the beginning of that corpus, that matched the size in tokenized English-side tokens of the 1.1 and 2.0

	Inuktitut→English			English→Inuktitut					
	BLEU			BLEU			YiSi-0		
Training Data/Model	Dev.	Dev.-test	Test	Dev.	Dev.-test	Test	Dev.	Dev.-test	Test
NH 1.1 (older) SMT	22.1	16.8	17.9	14.4	9.5	11.3	47.4	42.4	44.1
NH 1.1 (older) NMT	21.3	16.7	18.6	10.2	7.7	8.7	45.1	41.0	42.5
NH 1.1 (recent) SMT	31.9	24.9	26.0	21.6	14.4	16.5	53.1	47.7	49.1
NH 1.1 (recent) NMT	35.7	26.6	29.1	23.8	17.5	19.1	52.1	47.4	48.2
NH 2.0 (older) SMT	22.7	17.4	18.5	14.9	9.8	11.9	48.6	43.5	45.2
NH 2.0 (older) NMT	25.7	19.3	21.6	11.7	7.7	9.6	47.1	42.5	44.1
NH 2.0 (recent) SMT	32.7	25.3	27.1	22.8	15.3	17.5	53.7	48.6	49.9
NH 2.0 (recent) NMT	38.0	28.3	31.8	24.3	18.0	19.8	53.7	49.0	49.8
NH 3.0 beta SMT	33.4	25.7	27.4	23.6	16.0	18.4	54.1	48.6	50.4
NH 3.0 beta NMT	41.1	31.3	34.6	24.7	18.4	**20.3**	55.2	50.2	51.7
NH 3.0 SMT	33.5	25.9	27.6	23.2	15.8	18.0	54.4	48.8	50.4
NH 3.0 NMT	**41.4**	**31.6**	**35.0**	**25.0**	**18.5**	**20.3**	**55.6**	**50.6**	**51.8**
NH 3.0 syllabics SMT	34.1	26.0	27.8	22.9	15.4	17.4	54.1	48.7	50.3
NH 3.0 syllabics NMT	**41.4**	31.4	**35.0**	*24.2*	*17.9*	*19.3*	*55.6*	*50.6*	*51.9*

Table 5: Machine translation baseline experiment results reported in terms of lowercase word BLEU score for both directions and cased YiSi-0 score for into Inuktitut. Note that scores for the syllabics experiments into Inuktitut should not be compared directly to romanized results, as the test set preprocessing differs; the best scores for syllabics are italicized.

corpora; we call these **NH 1.1 (older)** and **NH 2.0 (older)**, respectively. As we would expect, for both NMT and SMT, the increase in data size improves translation quality in both translation directions. SMT only outperforms NMT on the NH 1.1 corpus size (approximately 4 million English tokens), while the NMT systems outperform SMT on the NH 2.0 size (approximately 6.7 million English tokens).[19]

The size comparison is not entirely fair on its own: it conflates data size and recency. To examine recency effects, we also selected subsets from the end of the NH 3.0 training corpus that matched the corresponding "**(older)**" subsets in size; we denote these "**(recent)**" in the table. For both NMT and SMT, we observe large (often 10+ BLEU) recency effects, with more recent data performing better.

BLEU has long been criticized for not correlating well enough with human judgments on translation quality (Koehn and Monz, 2006; Ma et al., 2019). This problem is more apparent when evaluating translation into morphologically complex languages because BLEU does not account for morphological and lexical variation between the reference and the MT output. Minor differences in morpheme choice are scored as badly as mistranslating a whole word. Thus, in addition to BLEU, we have evaluated the MT systems into Inuktitut using YiSi-0 (Lo, 2019), a word-level metric that incorporates character-level information, which has been shown to correlate better with human judgment than BLEU on translation quality into agglutinative languages, such as Finnish and Turkish, at both sentence level and document level (Ma et al., 2018; Ma et al., 2019). In general, the YiSi-0 results follow the same trends as BLEU: (1) better aligned training data improves translation quality; (2) larger training data improves translation quality; and (3) more recent training data improves translation quality. However, as we do not yet have human judgments on MT into Inuktitut, we urge caution in the interpretation of any automatic metric.

[19]For NMT, we use the settings found to be optimal for NH 3.0 throughout our remaining experiments; this may be suboptimal, as noted in Sennrich and Zhang (2019), but we still would not expect the smaller corpora to outperform the larger in this scenario.

5. Conclusion

The main contribution of the work described in this paper is the release of a corpus of approximately 1.3 million aligned Inuktitut–English sentence pairs drawn from the proceedings of the Legislative Assembly of Nunavut. Care was taken to ensure that the sentence alignment was as accurate as possible: the performance of different sentence alignment algorithms was compared to gold standard manual alignments performed by two experts fluent in Inuktitut and English. The aligned corpus generated by the best-performing algorithm is the one being released: NH 3.0. This corpus was used to carry out baseline SMT and NMT experiments, in which we observed general improvements based on increased data size, and improvements for NMT based on improved alignment quality.

This is, to our knowledge, the largest parallel corpus of a polysynthetic language ever released. We hope this corpus and paper will help spark interest in research on MT from and to polysynthetic languages, and help to answer many questions of scientific and practical interest.

6. Acknowledgements

This work would not have been possible without the cooperation of Riel Gallant, Legislative Librarian at the Legislative Library of Nunavut, John Quirke, Clerk of the Legislative Assembly of Nunavut, and Amanda Kuluguqtuq, Liz Aapak Fowler, and Gavin Nesbitt of Pirurvik Centre. We would also like to thank the 2019 Annual Jelinek Memorial Workshop on Speech and Language Technology (JSALT) for providing a venue for experiments and feedback on the beta version of this corpus. Finally, we wish to thank Benoît Farley for updating his Prosyl/Tunngavik text extractor to meet the needs of this project.

7. Bibliographical References

Ataman, D. and Federico, M. (2018). An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 97–110, Boston, MA, March. Association for Machine Translation in the Americas.

Ataman, D., Negri, M., Turchi, M., and Federico, M. (2017). Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *CoRR*, abs/1707.09879.

Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, page 81–89. Association for Computational Linguistics.

Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cherry, C. and Foster, G. F. (2012). Batch tuning strategies for statistical machine translation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 427–436. The Association for Computational Linguistics.

Cherry, C. (2013). Improved reordering for phrase-based translation using sparse features. In Lucy Vanderwende, et al., editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 22–31. The Association for Computational Linguistics.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R. M., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1370–1380. The Association for Computer Linguistics.

Ding, S., Renduchintala, A., and Duh, K. (2019). A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland, 19–23 August. European Association for Machine Translation.

Farley, B. (2009). Uqailaut. www.inuktitutcomputing.ca/Uqailaut/info.php.

Foster, G. (2010). 3-pass alignment approach in `ssal`. Portage software documentation.

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

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1352–1362. ACL.

Huang, L. and Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. In John A. Carroll, et al., editors, *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia, July. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Marti A. Hearst et al., editors, *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Larkin, S., Chen, B., Foster, G., Germann, U., Joanis, E., Johnson, H., and Kuhn, R. (2010). Lessons from NRC's Portage system at WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 127–132, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lepage, J.-F., Langlois, S., and Turcotte, M., (2019). *Evolution of the language situation in Nunavut, 2001 to 2016*. Statistics Canada, Ottawa, ON.

Li, X., Ge, N., Grimes, S., Strassel, S. M., and Maeda, K. (2010). Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., and Junker, M.-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Lo, C.-k. (2019). YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, August. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado, June. Association for Computational Linguistics.

Ma, Q., Bojar, O., and Graham, Y. (2018). Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 682–701, Belgium, Brussels, October. Association for Computational Linguistics.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. (2019). Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August. Association for Computational Linguistics.

Martin, J., Johnson, H., Farley, B., and Maclachlan, A. (2003). Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: Data driven machine translation and beyond, Volume 3*, pages 115–118. Association for Computational Linguistics.

Martin, J., Mihalcea, R., and Pedersen, T. (2005). Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74. Association for Computational Linguistics.

Micher, J. (2017). Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106. Association for Computational Linguistics.

Micher, J. (2018). Using the Nunavut Hansard data for experiments in morphological analysis and machine translation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 65–72, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, pages 135–144.

Nicholson, J., Cohn, T., and Baldwin, T. (2012). Evaluating a morphological analyser of Inuktitut. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 372–376, Montréal, Canada, June. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. (2005). Statistical machine translation: The fabulous present and future. Invited talk at the Workshop on Building and Using Parallel Texts, ACL 2005.

Okalik, P. (2011). Inuktitut and parliamentary terminology. *Canadian Parliamentary Review*, 34(4):22–24.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630, May.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Simard, M., Foster, G., and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of TMI-92*, pages 67–81, Montreal, Canada.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In John H. L. Hansen et al., editors, *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*. ISCA.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In *Pro-*

ceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1342–1348, Hong Kong, China, November. Association for Computational Linguistics.

Toral, A., Edman, L., Yeshmagambetova, G., and Spenader, J. (2019). Neural machine translation for English–Kazakh with morphological segmentation and synthetic data. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 386–392, Florence, Italy, August. Association for Computational Linguistics.

Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1387–1392. ACL.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yu, Q., Max, A., and Yvon, F. (2012). Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *The 5th Workshop on Building and Using Comparable Corpora*, page 10.