Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus

Starkaður Barkarson
The Árni Magnússon Institute
for Icelandic Studies
starkadur.barkarson@
arnastofnun.is

Steinþór Steingrímsson
The Árni Magnússon Institute
for Icelandic Studies
steinthor.steingrimsson@
arnastofnun.is

Abstract

We present ParIce, a new English-Icelandic parallel corpus. This is the first parallel corpus built for the purposes of language technology development and research for Icelandic, although some Icelandic texts can be found in various other multilingual parallel corpora. We map which Icelandic texts are available for these purposes, collect and filter aligned data, align other bilingual texts we acquired and describe the alignment and filtering processes. After filtering, our corpus includes 39 million Icelandic words in 3.5 million segment pairs. We estimate that our filtering process reduced the number of faulty segments in the corpus by more than 60% while only reducing the number of good alignments by approximately 9%.

1 Introduction

In recent years machine translation (MT) systems have achieved near human-level performance in a few languages. They rely heavily on large amounts of parallel sentences. This can pose problems for inflected languages like Icelandic, where a substantial amount of data is necessary to cover common word forms of frequent words. For training statistical (SMT) and neural (NMT) machine translation systems, parallel data quality is important and may weaken performance if inadequate, especially for NMT (see e.g. Khayrallah and Koehn (2018)). A vital part of compiling good parallel corpora is thus to assess how accurate the alignments are.

In addition to MT, parallel corpora have been employed for many tasks, including the creation of dictionaries and ontologies, multilingual and cross-lingual document classification and various annotation projection across languages. See e.g. Steinberger et al. (2012) for a discussion on the many aspects of parallel corpora usage.

This paper introduces ParIce, the first parallel corpus focusing only on the English-Icelandic language pair. There have been few available multilingual parallel corpora including Icelandic texts and those that exist vary in quality. Our primary aim was to build a corpus large enough and of good enough quality for training useful MT systems, while we also want it to be useful for other purposes, such as those listed above. The project plan for a language technology program for Icelandic, set to start in fall 2019, notes that for a quality MT system, a parallel corpus of 25-35 million sentence pairs is preferable, although 2 million may be sufficient for initial experiments with state-of-the-art methods (Nikulásdóttir et al., 2017). This first version of ParIce includes 3.5 million sentence pairs. That is quite far from the ambitious aim set forward in the project plan, but is hopefully sufficient to get meaningful results when used to train MT systems.

We started by mapping what parallel data was available and assessed its quality. We then collected unaligned bilingual texts and aligned and filtered them. In the filtering process we want to remove as many bad segment pairs as possible, while maximizing the number of good sentence pairs we hold on to. There is considerable literature on filtering parallel texts. Taghipour et al. (2011) point out that a lack of properly labeled training data makes it hard to use discriminative methods. They utilize unsupervised methods for outlier detection. To reduce reliance on labeled examples, Cui et al. (2013) conduct a PageRank-style random walk algorithm to iteratively compute the importance score of each sentence pair. The higher the score, the better the quality. Xu and Koehn (2017) tackle the problem of insufficient labeled data by creating synthetic noisy data

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 140–145
Turku, Finland, 30 September – 2 October,

to train a classifier that identifies known good sentence pairs from a noisy corpus.

In this paper we describe our semi-supervised method of using an NMT system trained on part of the corpus, and a bootstrapped dictionary to iteratively assess and score the sentence pairs. We then show how using the score to filter out low quality data results in a better quality corpus.

2 Available texts

The data mapping was twofold. First we looked for available parallel corpora with Icelandic and English texts. Then we looked for texts available to us in both languages that we could align and had permission to publish with open licenses.

2.1 Aligned data

We collected over 1.9 million English-Icelandic sentence pairs from other parallel corpora (see Table 1), mostly from the Opus project[1] but also from the Tilde MODEL corpus[2], ELRC[3] and a multilingual parallel bible corpus[4].

Opus (Tiedemann, 2012) has a variety of different parallel corpora in multiple languages. In the EN-IS language pair there are film and tv subtitles collected from OpenSubtitles[5], texts from localization files for KDE4, Ubuntu and Gnome and a collection of translated sentences from Tatoeba[6], an online collection of sentences and their translations, created by volunteers.

ELRC (European Language Resource Coordination) offers among others a parallel corpus that was derived from Icelandic and English texts from the Statistics Iceland (SI) website[7].

From the Tilde MODEL corpus (Rozis and Skadins, 2017) we include the EN-IS language pair from a corpus of texts from the European Medicines Agency document portal.

A parallel corpus of the bible in 100 languages (Christodoulopoulos and Steedman, 2015) is available online. This includes the Icelandic translation from 1981 and the King James version of the English bible.

Corpus	Sentence pairs	Bad alignments (%)
The Bible[4]	31,085	0.5
EMA[8]	420,297	3.3
Gnome[9]	5,431	n/a
KDE4[9]	87,575	45.0
OpenSubtitles[9]	1,368,170	8.3
Statistics Iceland[10]	2,360	8.0
Tatoeba[9]	8,139	0.0
Ubuntu[9]	2,127	2.5
TOTAL	**1,923,060**	

Table 1: Pair count and ratio of bad alignments in the parallel corpora available.

An examination of random sentences from these corpora revealed that the sentence pairs were sometimes faulty. This could be due to misalignment, mistranslation or other causes. Thus, in cases where we could obtain the raw data that the corpora were compiled from, we realigned them using our methods. For the EMA corpus we only had the raw data in pdf-files and decided against harvesting the texts from these files for realignment. The raw data for the SI corpus was not available on the ELRC-website, and we did not scrape the SI website for this project. The Tatoeba data is collected in such a way that there is no reason to align it again, and inspection of the data from GNOME indicated that the alignments were of insufficient quality and that mending them would prove hard so we decided to exclude them from our corpus.

2.2 Unaligned data

Regulations, directives and other documents translated for the members of the European Economic Area (EEA) were obtained from the EFTA-website[11], where they are available in both pdf and html format.

The Icelandic Sagas have been translated into numerous languages. Some of these translations are out of copyright and available in English on Project Gutenberg[12]. The Icelandic texts were obtained from the SAGA corpus (Rögnvaldsson and Helgadóttir, 2011). We also selected four books from Project Gutenberg, which were available in

[1] http://opus.nlpl.eu

[2] https://tilde-model.s3-eu-west-1.amazonaws.com/Tilde_MODEL_Corpus.html

[3] Available at the ELRC-SHARE repository

[4] http://christos-c.com/bible/

[5] http://www.opensubtitles.org

[6] http://tatoeba.org

[7] https://www.statice.is

[8] Part of the Tilde MODEL corpus

[9] Part of the OPUS corpus

[10] Created by ELRC

[11] https://www.efta.int

[12] https://www.gutenberg.org

translation on Rafbókavefurinn[13], a website with a collection of books in Icelandic in the public domain. The purpose of this was to experiment with aligning literary translations.

3 Compiling ParIce

We employed a two-step process to pair the sentences. First the texts were aligned with LF Aligner, except in cases where no alignment was necessary (see Section 3.1.2). Then the alignment was assessed and filtered.

3.1 Alignment

We used LF Aligner[14], which relies on Hunalign (Varga et al., 2005) for automatic sentence pairing. It aligns sentences in two languages by using a dictionary and information on sentence length.

3.1.1 Dictionary

We created a makeshift dictionary of over 12 thousand lemmas (D1) by scraping the Icelandic Wiktionary. In the case of nouns, pronouns and adjectives all possible inflections are listed on Wiktionary and were included in D1,

We ran LF Aligner, using D1, in the first pass of the alignment process. Afterwards the data was sent through the filtering process described in Section 3.2. We then used bitextor-builddics (Esplà-Gomis, 2009), to create another dictionary (D2). Builddics takes as an input source language segments in one file and target language segments in another. It then compares corresponding lines and builds a bilingual dictionary. Finally, D2 was expanded by getting all possible word forms of every Icelandic word in the dictionary from the Database of Modern Icelandic Inflection (DMII) (Bjarnadóttir, 2005). For each Icelandic word in D2, all the possible lemmas were found in DMII and every word form was retrieved for each lemma. D2 contains approx. 31 thousand lemmas, not counting different word forms. We used this dictionary for a second run of alignment on all the corpora.

3.1.2 Texts not requiring alignment

Texts from localization files (KDE4 and Ubuntu) are aligned by design. Some lines also contain strings that are not proper words but placeholders. In order to have less noisy texts these were removed. The Tatoeba segments were also not realigned, as the segments are created by translation.

3.2 Assessment and filtering

Aside from the cases where there was no need for assessment due to the nature of the data (Tatoeba, Bible), or because the alignments had already been filtered (KDE4, Ubuntu), we start by assessing the quality of the alignment and filtering out all lines deemed bad. A rough inspection of the aligned texts reveals that bad alignments usually come in chunks. If an error occurs in one alignment it has a tendency to affect the alignment of one or more sentences that follow since LF Aligner can take several lines to find the right path again.

As part of the filtering process we translate the English text of each sentence pair into Icelandic and then compare the translations to the Icelandic sentence and score each pair depending on that comparison. Every chunk of sentence pairs that has unfavorable scoring is deleted, but the sentence pairs that are not deleted are used to expand the dictionary (D2). These steps of translating, scoring, filtering and expanding the dictionary are repeated several times.

Before describing the filtering pipeline in detail we describe the scoring process.

3.3 Scoring

All English segments were translated into Icelandic by employing these two methods:

i) All possible translations were obtained from dictionary D2 for every word in the English sentence, thus creating a multiset for each word.

ii) We used OpenNMT (Klein et al., 2017) to train an MT system, using a 1 million segment translation memory provided by the Translation Centre of the Ministry for Foreign Affairs, and parallel corpora obtained from Opus. The system was used to translate each English sentence into Icelandic.

Since Icelandic is an inflected language it was necessary to take into account every word form. As described in Section 3.1.1, the D2 dictionary included all possible word forms but for the translated sentence obtained with OpenNMT all word forms were obtained by using DMII and a multiset created for each word of the translated sentence.

The score for every sentence was calculated by finding the average of score1 and score2. Score1 is the ratio of words in the Icelandic sentence found in any of the multisets created by either the first or

[13] https://rafbokavefur.is

[14] http://sourceforge.net/projects/aligner

	Before Filtering	Accepted Pairs	Accepted Pairs (%)	Bad Accepted Pairs (%)	Bad Deleted Pairs (%)
The Bible	32,964	32,964	100.0	0.0	n/a
Books	16,976	12,416	73.1	3.5	38.0
EEA	2,093,803	1,701,172	81.3	5.0	63.5
EMA	420,297	404,333	96.2	1.3	45.0
ESO	12,900	12,633	97.9	0.5	46.0
KDE4	137,724	49,912	36.2	9.0	n/a
OpenSubtitles	1,620,037	1,305,827	80.6	1.4	37.0
Sagas	43,113	17,597	40.8	11.0	55.5
Statistics Iceland	2,481	2,288	92.2	5.0	56.0
Tatoeba	8,263	8,263	100.0	0.0	n/a
Ubuntu	11,025	10,572	95.9	2.0	n/a
TOTAL	**4,399,582**	**3,557,977**	**80.9**		

Table 2: Pair count before and after filtering as well as ratio of accepted pairs and deleted pairs that were deemed bad during the assessment.

the second method of translation. Score2 was calculated by finding the ratio of multisets, created with dictionary D2, that contained a word form appearing in the Icelandic sentence, and the ratio of multisets, created with OpenNMT/DMII, that contained a word form appearing in the Icelandic sentence, and then selecting the higher ratio. Sentence pair (1) gets 1.0 as score1 since each word in the Icelandic sentence would be found in the multisets, and 0.38 as score2 since only three of eight multisets would contain a word appearing in the Icelandic sentence. The score would thus be 0.69.

(1) a. Hann gekk inn. *(e. He walked in)*
b. As he walked in he sang a song.

The score for each document is the average score for all sentence pairs.

3.3.1 Filtering

We set up a filtering pipeline, sending one subcorpus through it at a time. Steps 4-8 in the pipeline, detailed below, were repeated several times with the conditions for "good" sentence pairs strict at first but more lenient in later iterations. The conditions were controlled by thresholds and deletion rules described in step 7 below.

Our filtering pipeline is set up as follows:

1. Aligned sentences are cleaned of all out-of-vocabulary unicode symbols, as some symbols cause problems in parsing.

2. The aligned texts are divided into files, one for each document in the text. The process deletes faulty files, defining faulty to be ones that contain either unusually few and large aligned segments or a very low ratio of Icelandic letters (i.e. *ð, þ, ö*) in the Icelandic segments, indicating that they might have been obtained by inadequate OCR.

3. The English segments are automatically translated to Icelandic with the OpenNMT system, as described in Section 3.3.

4. The English segments are translated using dictionary D2, as described in Section 3.3.

5. Each sentence pair is scored.

6. Files receiving on average a score below a given threshold for their segment pairs are deleted. The assumption is that the English and Icelandic files being aligned are not compatible or only compatible in minor parts.

7. Sentences are deleted according to one of two rules: i) If a certain number of pairs in a row have a score under a given threshold, they are deleted. ii) If a certain number of pairs in a row have a score above a given threshold, they are not deleted but all other pairs are deleted. The second rule is more strict and is usually only used during the first iteration. For both rules, the number of pairs in a row and the threshold have different values selected for each iteration and subcorpus.

8. Two text files are created from the accepted sentence pairs: English in one file and Ice-

landic in the other, with sentences matching on line number. The files are used to create a new dictionary with bitextor-builddics which is appended to dictionary D2.

In OpenSubtitles there often exist many versions of both English and Icelandic subtitles for the same film. Therefore we sometimes chose between several files from the corpus. Working with the Sagas, we sometimes had two translations of the same Saga. The files receiving the highest score, after going through the pipe, were selected.

4 Resulting dataset

Before filtering the texts we had 4,399,582 sentence pairs in total, see Table 2. During the filtering process 841,605 pairs were deleted, 19.13%. The resulting dataset contains 3,557,977 pairs.

4.1 Quality assessment

We manually assessed the alignment quality of the new corpus as well as the pre-existing corpora by checking from 200 to 800 sentence pairs in each subcorpus, depending on its size. If the sentences were not in agreement, if a large chunk was erroneous or if a sentence in one language contained a segment not found in the other language the pair was classified as bad.

The quality of ParIce varies between subcorpora, from containing no bad alignments to 11.0%. Approximately 3.5% of the alignments in the corpus are bad, while the ratio was 8% in the pre-existing corpora. See Table 1 for quality estimates of the pre-existing corpora and Table 2 for quality estimates of ParIce.

5 Filter assessment

We checked a random sample of 100 to 400 of the deleted pairs in each subcorpus, depending on the number of deleted lines, and counted the amount of bad pairs. The results are shown in Table 2. When we compare this assessment to the assessment of the final version of ParIce, we can estimate the reduction of errors in the filtering process. If we exclude the alignments of KDE4 and Ubuntu, which were not sent through the main filtering pipeline, then 753,340 of 4,250,833 alignments, or 17.72%, were deleted during the filtering process. Of these 53.0% were bad, given that the ratio is the same as in our random samples, and the filtering process reduced the number of faulty segments in the corpus by 77.0% while it only reduced the number of good ones by 9.5%.

6 Availability

ParIce can be downloaded from `http://www.malfong.is`. Available sentences have been PoS-tagged with a BiLSTM tagger (Steingrímsson et al., 2019), lemmatized with Nefnir (Ingólfsdóttir et al., 2019) and word aligned with GIZA++ (Och and Ney, 2003).

The corpus is also searchable on `http://malheildir.arnastofnun.is` in a search tool powered by Korp (Borin et al., 2012).

7 Conclusion and future work

From a fragmented collection of around 1.9 million sentence pairs of unknown quality, and other data, we have built the ParIce corpus of approx. 3.5 million sentence pairs, assessed to be of acceptable quality. This enables the Icelandic language technology community, and others, to experiment with building MT systems for the English-Icelandic language pair.

While increasing alignment quality, our method filters out many perfectly good sentence pairs. It is necessary both to improve the filtering and the alignment processes. For better alignments a better dictionary is crucial. In the absence of a better dictionary, multiple iterations of aligning and filtering, where the aligned data is used to grow the dictionary in every iteration, could be helpful.

For better filtering adding features to our scoring algorithm might be beneficial. Hangya and Fraser (2018) follow Lample et al. (2018) and train monolingual word embeddings for two languages and map them to a shared space without any bilingual signal. They use these bilingual word embeddings for parallel corpus filtering. This approach could prove useful for our purposes.

The web is our best prospect for growing the corpus. We have yet to see how much the ParaCrawl project will collect of Icelandic parallel data, but can expect filtering to be important for that dataset (see e.g. Koehn et al. (2018)).

It would be useful to try to estimate how good MT systems trained on this data can get, and whether our filtering and realigning methods are useful for that purposes. Training MT systems on data from different stages and evaluating BLEU scores should thus be added as part of our pipeline, when working on future versions of ParIce.

References

Kristín Bjarnadóttir. 2005. Modern icelandic inflections. In H. Holmboe, editor, *Nordisk Sprogteknologi 2005*. Museum Tusculanums Forlag, Copenhagen, Denmark.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, pages 474–478.

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for SMT using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–345, Sofia, Bulgaria. Association for Computational Linguistics.

Miquel Esplà-Gomis. 2009. Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites. In *Proceedings of MT Summit XII*, Ottawa, Canada. Association for Machine Translation in the Americas.

Viktor Hangya and Alexander Fraser. 2018. An unsupervised system for parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 882–887, Belgium, Brussels. Association for Computational Linguistics.

Svanhvít Ingólfsdóttir, Hrafn Loftsson, Jón Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the* 22^{nd} *Nordic Conference of Computational Linguistics*, NODALIDA 2019, Turku, Finland.

Huda Khayrallah and Philipp Koehn. 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. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Anna Björk Nikulásdóttir, Jón Guðnason, and Steinþór Steingrímsson. 2017. *Language Technology for Icelandic 2018-2022: Project Plan*. Mennta og menningarmálaráðuneytið, Reykjavík, Iceland.

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

Roberts Rozis and Raivis Skadins. 2017. Tilde MODEL - multilingual open data for EU languages. In *NODALIDA*, pages 263–265. Association for Computational Linguistics.

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2011. Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change. In C. Sporleder, A. van den Bosch, and K. Zervanou, editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Springer, Berlin.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 454–459, Istanbul, Turkey. European Language Resources Association (ELRA).

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of RANLP 2019*, Varna, Bulgaria.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *MT Summit XIII. Machine Translation Summit (MT-Summit-11), 13., September 19-23, Xiamen, China*. NA.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, and Viktor Nagy Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.