Abstract

We present a new Icelandic–English parallel corpus, the Icelandic Parallel Abstracts Corpus (IPAC), composed of abstracts from student theses and dissertations. The texts were collected from the Skemman\textsuperscript{1} repository which keeps records of all theses, dissertations and final projects from students at Icelandic universities. The corpus was aligned based on sentence-level BLEU scores, in both translation directions, from NMT models using Bleualign. The result is a corpus of 64k sentence pairs from over 6 thousand parallel abstracts.

1 Introduction

Parallel text corpora are the cornerstone of machine translation systems. While recent developments have reduced this dependence somewhat with unsupervised neural machine translation they continue to play an important role, in particular during fine-tuning (Artetxe et al., 2019; Lample and Conneau, 2019).

Parallel data is also of high importance for automatic evaluation of machine translation models via computable metrics such as BLEU or NIST. Parallel corpora can also be used to automatically construct parallel glossaries or dictionaries.

Manual creation of parallel corpora is time consuming and expensive and naturally occurring texts are thus of great interest. For texts in Icelandic and English, one such source is the Skemman\textsuperscript{1} repository, which contains a collection of student theses and dissertations from all Icelandic universities, including some research papers from faculty. It has been hosted at the National Library of Iceland since 2008 and lists over 35,000 entries. In this work we gather all available files from the repository, locate and extract parallel abstracts and align the resulting segments.

1.1 Existing corpora

Most of the currently aligned parallel Icelandic–English data is found in the ParIce collection of parallel corpora (Barkarson and Steingrímsson, 2019). While extensive, the sub-corpus quality is either varying in quality or very domain-specific. Crowd-sourced datasets include OpenSubtitles and Tatoeba while the Icelandic sagas and Gutenberg literature often contain arcane language. The higher quality parallel data is mainly sourced from translated EEA-regulations, medicinal information (EMEA) or software localizations (Ubuntu and KDE). Other datasets contain vocabulary that may not be desirable such as religious texts (Jehova’s Witnesses corpus, JW300 (Agić and Vulić, 2019) and the Bible). It has therefore been difficult to automatically evaluate the broader generalization performance of existing translation models, something we hope to address with the wide scope and high quality of IPAC.

| Corpus                     | Size |
|---------------------------|------|
| The Bible                 | 33k  |
| EEA regulatory texts      | 1,700k |
| EMA                       | 404k |
| European Space Observatory (ESO) | 12.6k |
| OpenSubtitles             | 1,300k |
| Tatoeba                   | 10k  |
| Jehova’s Witnesses (JW300) | 527k |
| Other*                    | 93k  |
| IPAC (this work)          | 64k  |

Table 1: Existing parallel corpora, other* denotes software localizations, Project Gutenberg literature and the Icelandic sagas.

\textsuperscript{1}https://skemman.is

\textsuperscript{2}Checked on February 8, 2021.
Multilingual sources with more than 2 languages include Jehovah’s Witnesses corpus (JW300) and the European Medicines Association corpus (EMA/EMEA). For more information on the aforementioned corpora see (Tiedemann, 2012) and (Barkarson and Steingrimsson, 2019).

2 Abstract extraction

A small scraper was written in Python to download the PDF files belonging to each thesis entry. At the time of gathering a total of 31k PDF files were collected. This was necessary because not all entries included abstracts in their metadata, and only a small subset included abstracts in more than one language.

The repository provides a functionality for locking documents, optionally with a release date (potentially years in the future). Fortunately, most documents are not locked. Even so, many authors do not use this functionality at all and opt to encrypt their PDF with a password. While it seems the universities encourage or require abstracts in both English and Icelandic, not all documents include both regardless of the language of the document itself.

2.1 Language detection

The universities accept theses, dissertations and final projects in many languages, not just English and Icelandic. Unfortunately, the language of the main document is not part of the provided metadata, which only denotes the language of the provided abstract or title. A language might be listed in the keywords, but that was not a reliable indicator of the document language (especially so for Icelandic and English, which are usually implicit). Language detection based on abstract related keywords was used.

2.2 Text extraction

The text was extracted from the PDF files with the `pdftotext` software and various ad-hoc rules were written to determine the locations of the abstracts, such as via section title synonyms, length limits, lines starting on lowercase characters, comparable total lengths as well as the assumption that abstracts should occur near the beginning of a given file. A total of 7845 parallel abstracts were found.

3 Alignment

Sentence segmentation for Icelandic was performed with the Miðeind Tokenizer (Pörsteinsson et al., 2019) for Icelandic as well as English, as it was found to be accurate enough for both. NMT models were trained over a dataset composed of the pre-existing parallel corpora and backtranslated monolingual text from the news section of the Icelandic Gigaword Corpus (IGC) (Steingrimsson et al., 2018) and the English section of the newscrawl corpus. The models were then used to translate their respective source language side of the abstracts. Bleualign, an implementation of the algorithm described in (Sennrich and Volk, 2011), was then used to align the texts, leveraging the output of the NMT models in both translation directions.

| Lang.      | Segm. | Tokens pre | Tokens after |
|------------|-------|------------|--------------|
| Icelandic  | 84694 | 1656k      | 1324k        |
| English    | 83281 | 1811k      | 1483k        |
| Aligned    | 63870 | —          | —            |

Table 2: Alignment results

For a given document one or both translation directions can be used to compute alignments, however when both directions are provided the intersection of the alignments from each direction is used instead. The end result is a higher precision alignment, at the cost of lower recall.

| Field              | Abstracts | Acc. | Rej. |
|--------------------|-----------|------|------|
| Social sciences    | 2369      | 23962| 27502|
| Natural sciences   | 1248      | 11886| 13884|
| Medical and health | 1195      | 15546| 17388|
| Humanities         | 1026      | 10045| 11455|
| Business           | 604       | 5355 | 6095 |
| Mise               | 193       | 1723 | 1982 |

Table 3: Domain origin of parallel sentences

The grouping in Table 3 is based on the school or department where a given thesis was submitted. Note that the input here are individual sentences as opposed to the segments in Table 2 where

Part of Poppler, https://poppler.freedesktop.org/
some sentences may have been joined into a single many-to-one alignment.

3.1 Alignment quality

Due to the abstracts being written without any constraints of sentence-to-sentence translation some of them were found to align poorly due to content being omitted in either language. Fortunately most abstracts were almost translated at the sentence level and align well.

A random sample of 100 pairs was selected for manual evaluation and were classified into 4 categories: a) correct, b) near correct (slight loss of meaning, different choice of words), c) partial (some meaning completely lost, i.e. part of sentence gone or added). None were completely wrong, i.e. no alignment was present, as shown in Table 4.

| Group     | %   |
|-----------|-----|
| Correct   | 71% |
| Near correct | 22% |
| Partial   | 7%  |
| Incorrect | 0%  |

Table 4: Human evaluation

4 Discussion and future work

Most of the heuristics applied in the extraction and filtering stage were unnecessarily coarse for the sake of eliminating noise and increasing precision. More fine-grained heuristics and newer translation models may significantly increase the total yield without introducing much additional noise in future versions. Uncertainty estimation (Fomicheva et al. 2020) may also be able to identify poor alignments more accurately than any heuristics-based approach. BLEU is also a poor metric at the sentence-level and is typically used at the corpus level, a translation metric such as BERTScore (Zhang et al. 2020) which is a closer correlate with human judgement may help in this regard.

4.1 Release

The shuffled aligned parallel corpus is made available on the CLARIN-repository in pre-defined splits.\[7\]

5 Conclusion

We have extracted and aligned a high quality parallel Icelandic–English corpus IPAC from a wide variety of academic fields. It is orthogonal to other Icelandic–English parallel corpora and consists of a wide variety of topics. We envision it serves well, not only for training, but also as a much welcome benchmark of Icelandic–English machine translation systems and look forward to seeing it in use.

References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An Effective Approach to Unsupervised Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English–Icelandic Parallel Corpus. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, pages 140–145, Turku, Finland. Linköping University Electronic Press.

M. Fomicheva, Shuo Sun, L. Yankovskaya, F. Blain, Francisco Guzmán, M. Fishel, Nikolaos Aletrás, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. Transactions of the Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. In NeurIPS.

Vilhjálmur Porsteinsson, Hulda Óladóttir, and Hrafn Loftsson. 2019. A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In Proceedings - Natural Language Processing in a Deep Learning World, pages 1397–1404. Incoma Ltd., Shoumen, Bulgaria.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based Sentence Alignment of Parallel Texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), pages 175–182. Riga, Latvia. Northern European Association for Language Technology (NEALT).

Steinþór Steingrímsson, Sigrún Helgadóttir, Írirkur Rögnvaldsson, Starkaður Barkarson, and Jón
Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tianyi Zhang, V. Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *ICLR.*