Tilde MODEL - Multilingual Open Data for EU Languages

Roberts Rozis
Tilde
roberts.rozis@tilde.com

Raivis Skadiņš
Tilde
raivis.skadins@tilde.com

Abstract

This paper describes a Multilingual Open Data corpus for European languages that was built in scope of the MODEL project. We describe the approach chosen to select data sources, which data sources were used, how the source data was handled, what tools were used and what data was obtained in the result of the project. Obtained data quality is presented, and a summary of challenges and chosen solutions are described, too.

This paper may serve as a guide and reference in case someone might try to do something similar, as well as a guide to the new open data obtained.

1 Introduction

The European language technology community relies on public corpora such as the DGT-TM (Steinberger et al., 2012) and Europarl (Koehn, 2005) as a primary resource for developing machine translation (MT) and many other technologies for EU languages. DGT-TM is an invaluable asset, making it the most-viewed resource in the EU Open Data Portal. However, it is also very limited, covering only the legislative domain, therefore cannot lead to quality MT systems in other domains.

The lack of language resources is one of the biggest obstacles to the development of language technology in Europe. In an effort to overcome this obstacle, we have made a commitment to create new multilingual corpora for European languages – particularly the smaller languages that need them most – and make them openly available to researchers and developers.

As part of this initiative, we have identified and collected multilingual open data sets in multiple languages and several key domains. In addition, collected resources has been cleaned, aligned, and formatted using data-processing tools, thus rendering the corpora useable for developing new products and services.

At the end of the project (April, 2017), we plan to submit over 10M segments of multilingual open data for publication on the META-SHARE[1] repository, maintained by the Multilingual Europe Technology Alliance, and on the EU Open Data Portal[2]. These open data sets will be available to technology developers, researchers, localization companies, and machine translation providers. The corpora will provide a crucial resource for boosting the quality of MT engines, including the new breed of MT systems built with neural networks.

The activities have been undertaken as part of the ODINE Open Data Incubator for Europe[3], which aims to support the next generation of digital businesses and fast-track the development of new products and services.

2 Data Sources

In ODINE Tilde undertook to collect 10 million parallel segments of open parallel data from the Web.

The extensive approach would mean crawling hundreds of thousands of web sites in attempt to identify parallel content and trying to build a parallel corpus. The downside of this approach is immense crawling and computing requirements as well as our observation that only a small fraction of multilingual content is really parallel, mostly comparable or not really parallel at all. Going this way we would not fit in the time frame allocated (6 month) and would require intensive human work.

The intensive approach – first work on identifying sources and selecting web sites of public data containing large number of parallel texts. We chose this approach and selected domains, which have not been processed for at

[1] http://www.meta-share.org/
[2] https://data.europa.eu/euodp/en/data/
[3] https://opendataincubator.eu/

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 263–265,
Gothenburg, Sweden, 23-24 May 2017.

least for last 5 years. We ended up with the following data sources:

- RAPID – Press Releases database of European Commission[4] in all EU languages. The content of the press releases is translated precisely, which makes it interesting for being used as a source for parallel corpus.
- EMA – European Medicines Agency [5] documents - Descriptions of medicines and instructions of use of medicines as well as various medical conditions.
- Documents portal of European Economic and Social Committee[6]
- Web site of European Central Bank (ECB)[7]
- Web site of World Bank[8] – content on World Bank projects and activities in various regions of the world.

Corpora from EMA and ECB web sites have been collected before (Tiedemann, 2009), but a lot of new data have been published in these sites since that. Besides processing the mentioned major multilingual web sites, we also processed many small web sites to collect data in culture and travel domains; typical examples of such web sites are airbaltic.com, fold.lv, riga2014.org, umea2014.se, plzen2015.cz, wroclaw2016.pl, dss2016.eu etc.

In great degree, the data processing workflow is fixed and similar in top-level steps for any resource processed. When a data source candidate is selected, we crawl and download the data, convert it to a normalized format – plain text, and align the data resulting in data files usable for training MT systems – Moses format (parallel plain text files where file extension signifies the language of the file) and TMX format files. Each of these steps include many smaller steps and processes carried out depending on each and individual resource.

Crawling and downloading means exploring the structure how the data is held in the server and how we can reference each and individual page or file. Often this is a multi-step process where table of content pages must be crawled, or search by a list of keywords must be queried. Doing so we get a list of pages where the content can be found, or a starting page to that, or a list of files metadata. Only then we can get to the files or pages to download them into a local storage. Downloading hundreds of thousands of files/pages must be done politely in order not to abuse the remote server.

Conversion to a normalized format means extraction of plain text from whichever format be it PDF or HTML, or DOC, DOCX, etc. It is critical to retain the original text flow and structure and sequence of paragraphs. Last step of conversion is segmentation – the text is split in segments, mostly sentences which are the smallest granularity of data.

Alignment takes place by passing a large number of data, mostly aligned files of plain text segments to the alignment tool. The aligner builds statistical model of matching data, and selects matching segments, which are saved to the output.

3 Challenges & Solutions

Content downloading. We use custom-built and tailored PERL/Python scripts for downloading each of the selected resources. We do so to ensure usage of all the metadata (file identification by content and language) available from the source to obtain aligned files by language in the input. There are processing risks in other steps. The smaller languages are not so well represented in terms of parallel data, they tend to be more complex and contain inflections leading to potential alignment errors. We want to minimize that risk in this step; it is part of the approach chosen.

Content normalization. We used LibreOffice on Linux to convert various **office file formats** to plain text.

Dealing with **PDF** format files is still the hard nut. In rare and specific cases like with files of EMA (files originating from Microsoft Word) using commercial tools (Adobe Acrobat) yield in a very smooth process and high conversion quality. However, when working with content originating from QarkXPress or Adobe layout tools, use of a mix of other tools (Skadiņš et al., 2014) is needed. It remains a challenge to deal with

- Damaged files – files which might open for view but would take forever to process
- Protected DOCs/PDFs – we just have to skip those.
- Scanned PDFs – since multiple tools were used, we did not learn how to identify and skip those.

[4] http://europa.eu/rapid/
[5] http://www.ema.europa.eu/ema/
[6] https://dm.eesc.europa.eu/
[7] http://www.ecb.europa.eu/
[8] http://web.worldbank.org/

Segmentation. Until MODEL project we used our in-house as well as third party tools for segmentation of content before alignment. In this project we adapted SRX segmentation rules and integrated them with a Segment Program[9].

To perform **alignment of segments** we use Microsoft Bilingual Sentence Aligner[10] (Moore, 2002). We had to split the content in smaller packages, otherwise the aligner would stop working due to lack of memory.

Parallel computing. We did use the cloud computing potential to speed up processing of parallel data for multiple language pairs. With the use of Amazon Web Services cloud, we could achieve the result in 3 weeks for a resource that would otherwise take a whole year in a serial processing manner.

4 Results

We processed web sites described above and obtained the following data (Table 1) in the output. Corpus contains multilingual entries, English, French and German segments aligned with segments in all other represented languages. We selected random subsets of data and manually evaluated alignment quality, depending on language pair only 2-8% of segments contained alignment issues (Table 2).

	Files	**Langu-ages**	**Aligned segments**
RAPID	401K	24	4 M
EMA	81K	22	6 M
EESC portal	660K	36	6 M
ECB, World Bank	103K	32	70 K
Culture and travel domain	est. 5K	15	est. < 1M

Table 1: Amount of aligned data

	de-en	en-lv	en-pl
RAPID	92	96	93
EMA	95	92	99
EESC portal	99	92	96

Table 2: Human QA results (% of correctly aligned segments)

5 Conclusions

Scattered and isolated, without care about reuse in MT from the side of data origin, there does exist parallel content out there in the internet, which can be found, collected, processed and used in training MT systems. We are proud about being able to share with MT community over 16M of parallel segments of Tilde MODEL corpus collected and processed during MODEL project. At the end of the project (April, 2017), we released the corpus as open data and published on the META-SHARE[11] repository and on the EU Open Data Portal.

Not all data crawled has been processed and aligned, we had to discard huge amounts of PDF files due to unsatisfactory text extraction tools available. This leaves room for future work.

Acknowledgments

Work has been done in project "Multilingual Open Data for EU Languages (MODEL)" under the project framework "Open Data INcubator for Europe (ODINE ID: 644683)".

References

Koehn, P. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. In Conference Proceedings: the tenth Machine Translation Summit. Phuket, Thailand: AAMT, pp. 79-86

Moore, R.C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*. London, UK: Springer-Verlag, pp. 135-144.

Skadiņš R., Tiedemann J., Rozis R., Deksne D. 2014. *Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus*. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), pp. 1850–1855.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. 2012. *DGT-TM: A freely Available Translation Memory in 22 Languages*. Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012). Istanbul, Turkey, pp. 454-459.

Tiedemann, J. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pp. 237-248, John Benjamins, Amsterdam/Philadelphia

[9] https://github.com/loomchild/segment

[10] https://www.microsoft.com/en-us/download/details.aspx?id=52608

[11] http://metashare.tilde.com/repository/search/?q=Tilde+MODEL