iADAATPA Project: Pangeanic use cases

Mercedes García-Martínez, Amando Estela,
Laurent Bié, Alexander Helle, Manuel Herranz
m.garcia/a.estela/l.bie/a.helle/m.herranz@pangeanic.com

Abstract

The iADAATPA[1] project coded as N° 2016-EU-IA-0132 that ended in February 2019 is made for building of customized, domain-specific engines for public administrations from EU Member States. The consortium of the project decided to use neural machine translation at the beginning of the project. This represented a challenge for all involved, and the positive aspect is that all public administrations engaged in the iADAATPA project were able to try, test and use state-of-the-art neural technology with a high level of satisfaction.

One of the main challenges faced by all partners was data availability. Although all public administrations had some data available, it was clearly insufficient for high-level customization. In some cases, we had merely a few hundred words or several tens of thousand words. Each domain (field) has its own unique word distribution and neural machine translation systems are known to suffer a decrease in performance when data is out-of-domain.

Pangeanic is a language service provider (LSP) specialised in natural language processing and machine translation. It provides solutions to cognitive companies, institutions, translation professionals, and corporations. The problem faced by the iADAATPA project at Pangeanic was twofold:

1. Availability of training data in some language combinations.
2. How to successfully train a translation model on multi-domain data.

Language pairs and domains Pangeanic's use cases are for 2 Spanish public administrations: (1) Generalitat Valenciana (regional administration) translating from Spanish into and out of English, French, Catalan/Valencian, German, Italian, Russian and (2) SEGITTUR[2] (tourism administration) translating from Spanish into and out of English, French, German, Italian, Portuguese.

Data acquisition For translation from Spanish to Russian there was no available in-domain data. Therefore, 2 translators were contracted as part of the project to create 30,000 segments of in-domain data, translating public administrations websites. They also cleaned United Nations material and post-edited general-domain data that was previously filtered as in-domain following the "invitation model" (Hoang and Sima'an, 2014). For the other language pairs, the input material was 30,000 post-edited segments. The main part of the training corpora (approximately 75%) was part of Pangeanic's own repository harvested through web crawling and also OpenSubtitles (Tiedemann, 2012). The rest of the corpus was automatically validated synthetic material using general data from Leipzig (Goldhahn et al., 2012).

[1] http://iadaatpa.com/

[2] https://www.segittur.es/es/inicio/index.html

Engine customization The data was cleaned using the Bicleaner tool (Sánchez-Cartagena et al., 2018). The data was lowercased and extra embeddings were added in order to keep the case information. The tokenization used was the one provided by OpenNMT[3] and words were divided in subwords according to the BPE (Sennrich et al., 2016) approach. The models were trained with multi-domain data and we improved performance following a domain-mixing approach (Britz et al., 2017). The domain information was prepended with special tokens for each target sequence. The domain prediction was based only on the source as the extra token was added at target-side and there was no need for a-priori domain information. This approach allowed the model to improve the quality for each domain.

Acknowledgements The work reported in this paper was conducted during the iADAATPA project, which was funded by INEA through grant N° 2016-EU-IA-0132 as part of the EU's CEF Telecom Programme.

References

Britz, Denny, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126. Association for Computational Linguistics.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC12)*, pages 759–765.

Hoang, Cuong and Khalil Sima'an. 2014. Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928–1939. Dublin City University and Association for Computational Linguistics.

Sánchez-Cartagena, Víctor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 95–103, Brussels, Belgium, October. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May. European Languages Resources Association (ELRA).

[3] http://opennmt.net/