F. Soares, M. Krallinger

BVS CORPUS: A MULTILINGUAL PARALLEL CORPUS OF BIOMEDICAL SCIENTIFIC TEXTS AND TRANSLATION EXPERIMENTS

Abstract. The BVS database (Health Virtual Library) is a centralized source of biomedical information for Latin America and Carib, created in 1998 and coordinated by BIREME (Biblioteca Regional de Medicina) in agreement with the Pan American Health Organization (OPAS). Abstracts are available in English, Spanish, and Portuguese, with a subset in more than one language, thus being a possible source of parallel corpora. In this article, we present the development of parallel corpora from BVS in three languages: English, Portuguese, and Spanish. Sentences were automatically aligned using the Hunalign algorithm for EN/ES and EN/PT language pairs, and for a subset of trilingual articles also. We demonstrate the capabilities of our corpus by training a Neural Machine Translation (OpenNMT) system for each language pair, which outperformed related works on scientific biomedical articles. Sentence alignment was also manually evaluated, presenting an average 96 % of correctly aligned sentences across all languages. Our parallel corpus is freely available, with complementary information regarding article metadata.

Keywords. Parallel Corpora, Biomedical, Translation, Spanish, English.

1. Introduction

The availability of cross-language parallel corpora is one of the basis of current Statistical and Neural Machine Translation systems (SMT and NMT). Acquiring a high-quality parallel corpus that is large enough to train MT systems, specially NMT ones, is not a trivial task, since it usually demands human curating and correct alignment. In light of that, the automated creation of parallel corpora from freely available resources is extremely important in Natural Language Processing (NLP), enabling the development of accurate MT solutions. Many parallel corpora are already available, some with bilingual alignment, while others are multilingually aligned, with 3 or more languages, such as Europarl [Koehn 2005], from the European Parliament, JRC-Acquis [Steinberget et al. 2006], from the European Commission, OpenSubtitles [Zhang 2014], from movies subtitles.

The extraction of parallel sentences from scientific writing can be a valuable language resource for MT and other NLP tasks. The development of parallel corpora from scientific texts has been researched by several authors, aiming at translation of biomedical articles [Wu et al. 2011; Neves et al. 2016], or named entity recognition of biomedical concepts [Kors et al. 2015]. Regarding Portuguese/English and English/Spanish language pairs, the FAPESP corpus [Aziz and Specia 2011], from the Brazilian magazine revista pesquisa FAPESP, contains more than 150,000 aligned sentences per language pair, constituting an important language resource.

In Latin America and Carib, the Pan American Health Organization (OPAS), in agreement with BIREME (Biblioteca Regional de Medicina), maintains the BVS database, which is an important source of biomedical texts in three main languages: English, Spanish, and Portuguese. Currently, BVS has more than 1 million texts indexed, and also provides integrated search capabilities with PUBMED.

In this article, we explore the BVS database as a source of parallel corpora for the 3 aforementioned languages. We developed a trilingual parallel corpus with the 3 languages, as well as parallel corpora of English/Portuguese and English/Spanish abstracts. In addition, we provided various metadata regarding the publications.

2. Licensing

Most articles in the BVS database are open access documents. In order to avoid any copyright issues, we included in our datasets only open access documents. To retrieve license information, we crawled the BVS website containing information about the indexed journals[1] as well as the Directory of Open Access Journals[2.]

3. Materials and Methods

In this section, we detail the information retrieved from BVS website, the filtering process, the sentence alignment, and the evaluation experiments. Figure 1 shows the diagram of the steps followed for the development of the parallel corpora.

3.1 Document retrieval and parsing

BVS's website[3] offers simple and advanced search capabilities. We iteratively queried the database to retrieve all lists of results, which were then parsed and all relevant contents stored, such as authorship, title, and abstracts. We adopted the MongoDB database system, as it is document-oriented, and allows for the easy querying and storage of this type of data.

After the initial filtering, the resulting documents were processed for language checking[4] to make sure that there was no misplacing of abstract

[1] http://portal.revistas.bvs.br/

[2] https://doaj.org/

[3] http://bvsalud.org/

[4] https://github.com/Mimino666/langdetect

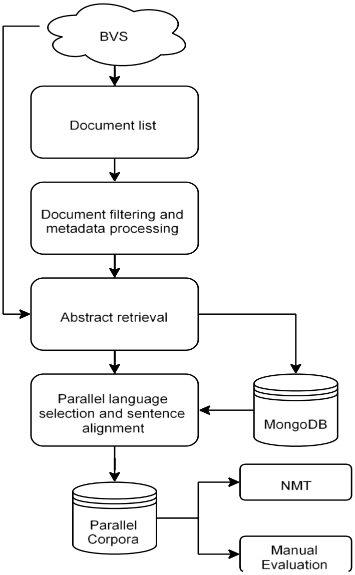
BVS
Document list
Document filtering and metadata processing
Abstract retrieval
Parallel language selection and sentence alignment
MongoDB
Parallel Corpora
NMT
Manual Evaluation

Fig. 1. **Method employed for corpora building**

language (e.g. English abstracts in the Portuguese field, or the other way around), removing the documents that presented such inconsistency. In addition, we also removed newline/carriage return characters (i.e \n and \r), as they would interfere with the sentence alignment tool.

3.2 Sentence alignment

For sentence alignment, we used the LF aligner tool[5], a wrapper around the Hunalign algorithm [Varga et al. 2005], which provides an easy to use and complete solution for sentence alignment, including pre-loaded dictionaries for several languages.

Hunalign uses Gale-Church sentence-length information to first automatically build a dictionary based on this alignment. Once the dictionary is built, the algorithm realigns the input text in a second iteration, this time combining sentence length information with the dictionary. When a dictionary is supplied to the algorithm, the first step is skipped. A drawback of Hunalign is that it is not designed to handle large corpora (above 10 thousand sentences), causing large memory consumption. In these cases, the

[5] https://sourceforge.net/projects/aligner/

algorithm cuts the large corpus in smaller manageable chunks, which may affect dictionary building.

The parallel abstracts were supplied to the aligner, which performed sentence segmentation followed by sentence alignment. After sentence alignment, the following post-processing steps were performed: (i) removal of all non-aligned sentences; (ii) removal of all sentences with fewer than three characters, since they are likely to be noise.

3.3 Machine translation evaluation

To evaluate the usefulness of our corpus for MT purposes, we trained an NMT model using the OpenNMT system [Klein et al. 2017] for all language pairs. The produced translations were evaluated according to the BLEU score [Papineni et al. 2002].

3.4 Manual evaluation

Although the Hunalign algorithm usually presents a good alignment between sentences, we also conducted a manual validation to evaluate the quality of the aligned sentences. We randomly selected 300 sentences, 100 for the triligual subset, and 100 for each subset of EN/PT and EN/ES. If the pair was fully aligned, we marked it as ”correct”; if the pair was incompletely aligned, due to segmentation errors, for instance, we marked it as ”partial”; otherwise, when the pair was incorrectly aligned, we marked it as ”no alignment”.

4. Results and Discussion

In this section, we present the corpus’ statistics and quality evaluation regarding NMT system, as well as the manual evaluation of sentence alignment.

4.1 Corpus statistics

Table 1 shows the statistics (i.e. number of sentences) for the aligned corpus according to the 2 language pairs and the trilingual subset. The dataset is available[6]in TMX format [Rawat et al. 2016], since it is the standard format for translation memories. We also made available the aligned corpus in an SQLite database in order to facilitate future subset selection. In this database, we included the following metadata information: year, keywords

[6] 10.6084/m9.figshare.8067533

in the available languages, database of origin, country, authorship, and URL for the full-text when available.

Table 1. Corpus **statistics according to language pair**

Language Pairs	Sentences
EN/PT	711,475
EN/ES	789,547
EN/PT/ES	203,719

4.2 Translation experiments

Prior to MT experiments, sentences were randomly split in three disjoint datasets: training, development, and test. Approximately 14,000 sentences were allocated in the development and test sets, while the remaining was used for training. For the NMT experiment, we used the Torch implementation[7] to train a 2-layer LSTM model with 500 hidden units in both encoder and decoder, with 20 epochs. During translation, the option to replace UNK words by the word in the input language was used.

Table 2 presents the BLEU scores for both translation directions with the 3 language pairs for the development and test partitions. We also included the best scores from a similar parallel corpus from Scielo [5] as a benchmark.

Table 2. **BLEU scores for translation using OpenNMT for the development and test partitions. Previous related work by Neves et al.(2016) is also presented for comparison in the right-hand column as benchmarking**

Language Pairs		Dev	Test	Bench
EN-ES	EN→ES	34.80	34.96	32.75
	ES→EN	33.82	34.28	30.53
PT-ES	PT→ES	55.78	56.11	—
	ES→PT	56.26	56.50	—
EN-PT	EN→PT	35.62	36.03	33.37
	PT→EN	35.88	36.12	31.78

[7] http://opennmt.net/OpenNMT/

Our models achieved better performance than the benchmark for all language pairs and directions, with at least 2.21 percentage points (pp) higher for the EN/ES language pair, achieving a maximum of 4.34 pp for the EN/PT language pair. It is noticeable the high scores achieved in the ES/PT pair, which we expect to be due to the high similarity between both languages.

4.3 Sentence alignment quality

We manually validated the alignment quality for 300 sentences randomly selected from the parsed corpus and assigned quality labels according Section 3.4. From all the evaluated sentences, average 96 % were correctly aligned, while average 2 % were partially aligned. The trilingual subset was the one with the best alignment, achieving 97 % correct alignment. The small percentage of no alignment is probably due to the use of Hunalign algorithm with the provided dictionaries. Figure 2 shows the alignment accuracy for all language subsets.

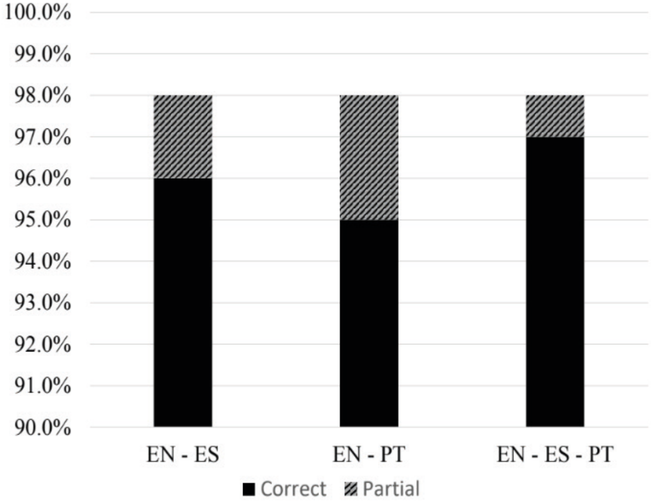

Fig. 2. **Alignment accuracy for the three language subsets**

5. Conclusion and future work

We developed a parallel corpus of biomedical abstracts in English, Spanish, and Portuguese. Our corpus is based on the BVS database, which contains biomedical texts from several sources in Latin America and Carib. The corpus contains the EN/ES, EN/PT language pairs as well as a trilingual subset of EN/PT/ES sentences.

Our corpora were evaluated through NMT experiments with OpenNMT system, presenting superior performance regarding BLEU score than a related work with a similar corpus. The NMT model presented remarkable results for the PT/ES language pair, possibly due to the similarity between the languages. We also manually evaluated sentences regarding alignment quality, with average 96 % of sentences correctly aligned.

For future work, we foresee the use of the presented corpus in mono and cross-language text mining tasks, such as text classification and clustering. As we included several metadata, these tasks can be facilitated. Other machine translation approaches can also be tested, including the concatenation of this corpus with other multi-domain ones.

Acknowledgments

This work was supported by the Encargo de Gestion SEAD-BSC of the Spanish National Plan for the Advancement of Language technologies, the ICTUSnet INTERREG Sudoe programme, the European Union Horizon2020 eTransafe (grant agreemed 777365) project, and the Amazon AWS Cloud Credits for Research.

References

1. Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5, pages 79–86, 2005.
2. Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. arXiv preprint cs/0609058, 2006.
3. Shikun Zhang, Wang Ling, and Chris Dyer. Dual subtitles as parallel corpora. 2014.
4. Cuijun Wu, Fei Xia, Louise Deleger, and Imre Solti. Statistical machine translation for biomedical text: are we there yet? In AMIA Annual Symposium Proceedings, volume 2011, page 1290. American Medical Informatics Association, 2011.
5. Mariana Neves, Antonio Jimeno Yepes, and Aurelie N´ ev´ eol. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may 2016. European Language Resources Association (ELRA).
6. Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. A multilingual gold-standard corpus for biomedical concept

recognition: the mantra gsc. Journal of the American Medical Informatics Association, 22(5):948–956, 2015.

7. Wilker Aziz and Lucia Specia. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In STIL 2011, Cuiaba, MT, October 2011.´
8. Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron. Parallel corpora for medium density languages. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE, page 247, 2005.
9. G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: OpenSource Toolkit for Neural Machine Translation. ArXiv e-prints, 2017.
10. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings *of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
11. Sunita Rawat, M. B. Chandak, and Nekita Chauhan. *An Approach for Efficient* Machine *Translation Using Translation Memory*, pages 285–291. Springer Singapore, Singapore, 2016.

Felipe Soares
Barcelona Supercomputing Center (BSC) — Spain
E-mail: felipe.soares@bsc.es
Martin Krallinger
Barcelona Supercomputing Center (BSC) — Spain
E-mail: martin.krallinger@bsc.es