Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3769–3774
Marseille, 11–16 May 2020

ParaPat: The Multi-Million Sentences Parallel Corpus of Patents Abstracts

Felipe Soares[1,2], Mark Stevenson[1], Diego Bartolome[2] and Anna Zaretskaya[2]
(1) University of Sheffield, Computer Science Department, 211 Portobello Road, Sheffield, S1 4DP, UK
(2) TransPerfect Translations, Passeig de Gràcia, 11, Barcelona, Spain
fs@felipesoares.net, mark.stevenson@sheffield.ac.uk, dbartolome@translations.com, azaretskaya@translations.com

Abstract

The Google Patents is one of the main important sources of patents information. A striking characteristic is that many of its abstracts are presented in more than one language, thus making it a potential source of parallel corpora. This article presents the development of a parallel corpus from the open access Google Patents dataset in 74 language pairs, comprising more than 68 million sentences and 800 million tokens. Sentences were automatically aligned using the Hunalign algorithm for the largest 22 language pairs, while the others were abstract (i.e. paragraph) aligned. We demonstrate the capabilities of our corpus by training Neural Machine Translation (NMT) models for the main 9 language pairs, with a total of 18 models. Our parallel corpus is freely available in TSV format and with a SQLite database, with complementary information regarding patent metadata.

Keywords: Parallel Corpus, Patents, Machine Translation Models

1. Introduction

The availability of parallel corpora is required by current Statistical and Neural Machine Translation systems (SMT and NMT). Acquiring a high-quality parallel corpus that is large enough to train MT systems, particularly NMT ones, is not a trivial task due to the need for correct alignment and, in many cases, human curation. In this context, the automated creation of parallel corpora from freely available resources is extremely important in Natural Language Processing (NLP). Many parallel corpora are already available, some bilingually aligned (Tiedemann, 2012) and others multilingually aligned, with 3 or more languages (e.g. Europarl from the European Parliament (Koehn, 2005), JRC-Acquis from the European Commission (Steinberger et al., 2006) and OpenSubtitles consisting of film subtitles (Zhang et al., 2014)).

The extraction of parallel sentences from patents and scientific texts can be a valuable language resource for MT and other NLP tasks. The problem has been researched by several authors to support, for example, translation of biomedical articles (Wu et al., 2011; Neves et al., 2016) and named entity recognition of biomedical concepts (Kors et al., 2015).

The most comprehensive dataset released to date for multilingual patent analytics and machine translation is the COPPA parallel corpus developed by the World Intellectual Property Office (WIPO)[1]. Version 2 of the COPPA corpus contains almost 13 million parallel sentences in 8 language pairs. However, all language pairs are from/to English and highly imbalanced, with the English/French language pair consisting of around 10 million sentences, more than 80% of the entire corpus. In addition, this corpus is copyrighted, being free only for research purposes and requiring an application process in order to gain access.

This work describes the development of a parallel corpus of patents abstracts from the Google Patents dataset. The corpus contains more than 68 million parallel sentences from 22 language pairs, and additional 96,000 parallel documents for other 52 language pairs, totalling 74 language pairs. An abstract aligned version of the corpus is available and can be used for purposes such as parallel sentence filtering and corpus linguistics analysis.

The main differences with regard to existing WIPO corpus are: (i) our corpus contains more than 70 language pairs, of those 9 language pairs are suitable for MT, since they have more than 100,000 sentences, (ii) the corpus is licensed under Creative Commons v4.0, (iii) both the parallel sentences and the metadata regarding the patents are included, allowing researchers to carry out a range of NLP tasks, such as classification and similarity evaluation and (iv) in addition to the parallel sentences, non-aligned sentences have also been made available, allowing researchers to view this corpus as a comparable one.

2. Licensing

The contents of the Google Patents database are licensed under the Creative Commons Attribution 4.0 International (CC-BY-4.0). This license allows one to use, share, remix and make commercial usage of the data as long as the original provider/author is given credit for their work (Commons, 2013).

To be sure we are adhering to the license, and to provide researchers with additional data, the patent ID is included as as part of the metadata.

3. Material and Methods

This section details the information gathered from Google, the filtering process and the method for abstract parsing and alignment.

3.1. Data retrieval

Google makes patents data available under the Google Cloud Public Datasets[2]. The data is accessible using the BigQuery platform in the following

[1] https://www.wipo.int/export/sites/www/patentscope/en/data/pdf/wipo-coppa-technicalDocumentation.pdf

[2] https://cloud.google.com/public-datasets/

website `https://console.cloud.google.com/bigquery?p=patents-public-data&d=patents&page=dataset`.

BigQuery is a Google service that supports the efficient storage and querying of massive datasets which are usually a challenging task for usual SQL databases. For instance, filtering the September 2019 release of the dataset, which contains more than 119 million rows, can take less than 1 minute for text fields. The on-demand billing for BigQuery is based on the amount of data processed by each query run, thus for a single query that performs a full-scan, the cost can be over USD 15.00, since the cost per TB is currently USD 5.00.

To retrieve the whole database, we accessed the BigQuery Patents repository and selected the last available release (from 16-September-2019). The whole 119 million rows in JSON format amounts to around 1.76 TB of data. We downloaded the database in JSON format and stored it locally for further processing. We could instead have used the capabilities of BigQuery, but it would have been expensive since several queries would be needed to create the parallel data.

3.2. Data parsing and patent alignment

The data from BigQuery was downloaded in chunks of newline-delimited JSONs, resulting in 2002 files. Data from BigQuery contain several attributes that may not be of interest when building a parallel corpus, such as prior-work and citing patents. Since only the abstracts are available in several languages (and not claims or full-text), we restricted our approach to the abstracts. The following steps describe the process of producing patent aligned abstracts:

1. Load the n^{th} individual file
2. Remove rows where the number of abstracts with more than one language is less than 2 for a given *family_id*. The family_id attribute is used to group patents that refers to the same invention. By removing these rows, we remove abstracts that are available only in one language.
3. From the resulting set, create all possible parallel abstracts from the available languages. For instance, an abstract may be available in English, French and German, thus, the possible language pairs are English/French, English/German, and French/German.
4. Store the parallel patents into an SQL database for easier future handling and sampling.

3.3. Sentence alignment

Dictionaries provided by the LF aligner tool[3] for sentence alignment are used to run the Hunalign algorithm (Varga et al., 2005), which provides an easy to use and complete solution for sentence alignment.

Parallel abstracts were extracted from the database for each language pair with more than 1,000 aligned abstracts and batch processing files created for use with Hunalign. Amazon EC2 r5dn.8xlarge instances, featuring 32 cores and 256GiB of RAM memory were used for this alignment step. All processing was done in-memory in order to avoid the disk writing overhead. Only final aligned files were then written to an EBS SSD drive.

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; (iii) removal of HTML tags, such as <p>, </p>, which we identified as present in some abstracts.

3.4. Machine translation evaluation

To evaluate the usefulness of our corpus for NMT purposes, we used it to train an automatic translator with FairSeq (Ott et al., 2019). The produced translations were evaluated according to the BLEU score (Papineni et al., 2002), using the evaluation tool SacreBLEU (Post, 2018).

4. Results and Discussion

In this section, we present statistics about the corpus and a quality evaluation in terms of BLEU scores using NMT.

4.1. Corpus statistics

Table 1 shows the number of parallel abstracts for all 44 language pairs included in our dataset. From more than 40 million abstracts, English/Chinese, English/Japanese, and English/French account for more than 80% of the total number of parallel abstracts, with English/Chinese being the one with most information. At this point, no sentence alignment has been performed, leaving the corpus aligned by abstract, thus it can be used for text analytics purposes or corpus linguistics approaches for translation equivalents or terminology extraction. We only selected the top 9 language pairs in terms of number of sentences to perform MT experiments, since the other ones are not large enough for NMT training. We opted to make available sentence aligned corpora for the top 23 languages, since they account for more than 9,000 abstracts.

The datasets are available online[4] in TSV format for the parallel abstracts and in Moses format for the parallel sentences (i.e. one sentence per line). Besides the aligned abstracts and sentences, we included the family_id for each abstract, such that researchers can use that information for other text mining purposes.

4.2. NMT experiments

An additional rule based filtering step to remove segments that might be misaligned was carried out prior to the NMT experiments. The source/target length ratio for each sentence was computed and those deviating too far from the average numbers in OPUS were removed (Tiedemann, 2012). In the following step, all sentences were randomly split in three disjoint datasets for each language pair: training, tuning and test. Test and tuning sets consist of 5,000 and 10,000 sentences, respectively, while the remainder of the sentences were used for training. As for pre-processing,

[3] `https://sourceforge.net/projects/aligner/`

[4] `https://github.com/soares-f/parapat`

Table 1: Corpus statistics regarding number of parallel abstracts in the corpus without any sentence alignment or preprocessing.

Language Pair	Abstracts
EN/ZH	18,480,037
EN/JA	8,154,020
EN/FR	6,413,137
EN/KO	2,228,868
DE/EN	1,588,458
EN/RU	1,364,068
DE/FR	704,914
FR/JA	498,367
EN/ES	360,638
FR/ZH	201,373
FR/KO	120,607
EN/UK	89,227
RU/UK	85,963
CS/EN	78,978
EN/RO	48,789
EN/HU	42,629
ES/FR	32,553
EN/SK	23,410
EN/PT	23,122
BG/EN	16,177
FR/RU	10,889
EL/EN	10,855
EN/IT	9,618
JA/ZH	7,924
RO/RU	7,663
EN/SL	6,423
KO/ZH	6,216
FR/PT	6,004
EN/NL	5,247
JA/KO	5,196
DE/ZH	3,745
EN/LT	3,594
EN/SR	3,517
EN/SH	3,062
AR/EN	2,360
ES/ZH	2,181
AR/FR	1,954
PT/ZH	1,835
DE/JA	1,780
ES/JA	1,718
EN/LV	1,631
ES/KO	1,375
DE/ES	1,175
EN/FI	1,173

sentences were tokenized with SentencePiece[5] with a non-shared vocabulary size of 32,000. The translation models were built using FairSeq with the transformer architecture. Hyperparameters are listed below:

- arch = transformer
- max-taget-positions = 64

Table 2: Size of the training corpora for the 9 language pairs in which NMT models were trained.

Language Pair	# Sentences	# Tokens (source)
EN/ZH	4.9M	155.8M
EN/JA	6.1M	189.6M
EN/FR	12.2M	455M
EN/KO	2.3M	91.4M
EN/DE	2.2M	81.7M
EN/RU	4.3M	107.3M
DE/FR	1.2M	38.8M
FR/JA	0.3M	9.9M
EN/ES	0.6M	24.6M

- min-lr = 1e-09
- label-smoothing = 0.1
- update-freq = 1
- warmup-init-lr = 1e-07
- dropout = 0.3
- weigth-decay = 0.9
- input_shapes = 128x64

Training was performed using Google Cloud TPUs v2. For all trained models we release the SentencePiece and the translation models, as well as the data already split into training/tuning/test for the sake of reproducibility and the possible usage of our developed corpora as a baseline for further improvement.

Table 2 shows the final number of parallel sentences used for training the NMT systems. For Asian languages, the number of sentences is greatly reduced when compared with Table 1, see later discussion.

Table 3 presents the BLEU scores for each language pair for the test sets. One can see that European languages present much higher BLEU scores than other language pairs. This is common phenomen in MT when languages share a similar grammar construction or lexicon. An interesting finding is regarding the BLEU points for English/Spanish, which are much higher than English/German, even with a corpus 3 times smaller. This has already been reported in other studies, such as Europarl(Koehn, 2005), with Spanish being the easiest language to be translated into and also achieving one of the best scores, just behind English/French. As for Asian languages, Korean/English presented the best scores, while English/Chinese language pair achieved discouraging results. We suspect this is due to possible misalignment and/or translation divergences.

4.2.1. Note on Asian Languages

BLEU scores between European and Asian languages are lower than between European languages. Hsu (2014) pointed out that translation divergences can affect MT output. According to Barnett et al. (1991), even though a set of legal translations can be valid, there is the notion of preferred translation that is not easily defined, but can be seen

[5] `https://github.com/google/sentencepiece`

Table 3: BLEU scores for translation using FairSeq in the test sets.

Language Pair	**Test**
ES→EN	57.71
FR→EN	49.60
EN→RU	49.10
EN→FR	48.39
DE→EN	48.35
EN→DE	46.77
RU→EN	46.73
EN→ES	44.98
DE→FR	42.13
FR→DE	36.56
KO→EN	23.23
JA→EN	17.78
EN→JA	13.15
ZH→EN	13.01
EN→KO	11.93
EN→ZH	10.29
JA→FR	9.52
FR→JA	7.10

when the natural translation differs from the source in a significant way. Based on the works of Lin et al. (2005) and Barnett et al. (1991), we identified the main translation divergences in patents between English and Asian languages (i.e. Korean, Japanese, and Chinese) as thematic and discourse. Below we give examples of both in the Chinese language.

- Thematic: when the arguments appears in different thematic roles, such as changing the focus of the sentence or changing from passive to active voice.

  > Natural English: *A method and system for recovering video monitoring service are disclosed in the present invention, which belong to video monitoring field.*
  > Natural Chinese: 本发明公开了一种视频监控业务恢复的方法和系统，属于视频监控领域.
  > Literal English: *The invention discloses a method and a system for recovering a video monitoring service, and belongs to the field of video monitoring.*

- Discourse: this happens when the difference between source and target texts is too large to be accounted for by other local transformations. In some cases, the whole structure of the text will be different in the two languages. This may happen also when localizing the translation to a specific purpose/country.

  > Natural English: *The gain factor of the E-DPDCH in the compressed mode is determined according to the code channel number which is required when transmitting data initially in the embodiments of the present invention, so it can be realized that the gain factor of the E-DPDCH in the compressed mode is determined accurately, furthermore the transmission power of the E-DPDCH is determined according to the gain factor, the waste of the transmission power of the E-DPDCH is reduced [...]*
  > Natural Chinese: 本发明实施例根据数据初次传输时所需的码道数，确定压缩模式下E-DPDCH的增益因子，实现了准确确定压缩模式下E-DPDCH的增益因子，进而根据该增益因子确定E-DPDCH的发射功率，减少了E-DPDCH的发射功率浪费.[...]
  > Literal English: *In the embodiment of the present invention, the gain factor of E-DPDCH in the compression mode is determined according to the code channel number which is required for the first transmission of data, and the gain factor of E-DPDCH in the compression mode is accurately determined, and then the transmission power of the E-DPDCH is determined according to the gain factor, reducing the emission power waste of E-DPDCH[...]*

Our linguists at TransPerfect Translations also identified that some patents that were presumably written first in Chinese (i.e. were initially submitted to the Chinese Patent Office) are not written in fluent Chinese, use non-conventional wording and/or lack cohesion. An example is given below, where the bold segments were flagged as not fluent with the sentences around them. The final bold segment conveys the same meaning as the first sentence of the paragraph but is connected by a comma to an unrelated segment, leading to a confusing overall paragraph.

> 本实用新型公开了汽车检测装置技术领域的一种汽车后桥减速器检测装置，**包括装置底座**，装置底座的顶部固设有龙门安装架，龙门安装架的内腔左右两侧均固设有升降滑轨和升降滑块，龙门安装架的顶部中心处固设有驱动电机、升降丝杆和升降套筒，升降套筒的左右两侧壁与左右两侧升降滑块之间均分别固接有安装连杆，安装连杆的前侧壁中部通过U形安装座连接有支撑悬臂梁和限位凹槽，限位凹槽的正上方设置有弧形压紧板，弧形压紧板的底部左侧与支撑悬臂梁之间铰接有压紧支杆，弧形压紧板的底部右侧铰接有锁紧螺杆，支撑悬臂梁的末端开设有螺杆卡槽，锁紧螺杆的外壁螺接有锁紧螺母，**本实用新型提供了一种汽车后桥减速器检测装置**。

In other cases, even though it is not stated in the Google Patents website, the available translations in English are very likely to be provided by MT, since they lack cohesion and fluency. In light of that, we suggest caution when using this language pair.

5. Comparison to other resources

The most comparable resource to the one presented here is the COPPA Corpus version 2 which contains around 13

Table 4: Comparison of some of our corpus language pairs to the language pairs present in the COPPA corpus V2 regarding number of documents and sentences.

Language Pair	Docs COPPA	Sents COPPA	Docs ParaPat	Sents ParaPat
EN/FR	2,570,292	10,557,032	6,413,137	18,360,221
EN/ZH	83,359	195,317	18,480,037	17,599,236
EN/JA	312,664	1,036,614	8,154,020	16,621,979
EN/RU	6,972	37,261	1,364,068	6,057,707
EN/KO	41,093	120,534	2,228,868	3,200,906
EN/DE	289,287	982,510	1,588,458	2,994,162
EN/ES	18,303	62,057	360,638	818,044
EN/PT	2,001	7,000	23,122	80,501
Total	3,321,970	12,991,325	38,589,226	65,732,756

million sentences. Table 4 compares the COPPA corpus version 2 to our corpus in terms of number of documents and sentences. We can see that for some language pairs, such as Chinese/English and Russian/English, our corpus is, respectively, 90 and 162 times larger in terms of number of sentences. For other language pairs the size comparison is more modest but still considerable, for French/English and German/English it contains 1.7 and 3 times more sentences. On average across all language our corpus contains 5 times the number of sentences as the COPPA corpus.

The Opus project (Tiedemann, 2012) is a database of open source parallel corpora. The collection is continuously growing as new corpora are made available. The Opus project contains more sentences than our corpus for some language pairs, e.g. for Chinese/English our corpus contains around 17.6M sentences and the UN corpus (Rafalovitch et al., 2009) 19.9M. However, for other language pairs our corpus is larger, e.g. for there are 6.6M English/Japanese sentences while the JW300 corpus (Agić and Vulić, 2019) contains around 2.1M. The UN corpus consists of translations of the United Nations general assembly resolutions and the JW300 mainly from the Jehovah's Witnesses website. However, our corpus is completely focused on the single domain of patents.

6. Conclusion and Future Work

We developed a parallel corpus of Patents abstracts in 74 language pairs, with 22 of them being also sentence aligned and the remaining 52 aligned at the abstract level. Additionally to the language pairs, we also provided 18 translation models based on Transformers and trained with FairSeq for the 9 largest language pairs. Our corpus is based on the patents abstracts database released by google, which is available under open-access license (CC-BY 4.0), thus favoring distribution and modifications.

We evaluated our corpus with an NMT experiment with Transformer models in FairSeq. Our translation experiment showed that the developed corpus is adequate for NMT purposes and provide a significant increase in size when compared to COPPA v2. We highlight the high translation scores achieved for European language pairs, boosted by the large number of sentences in English/French and English/German. Other important features of our corpus are the availability of parallel abstracts for low-resource languages for this domain, such as Arabic/French Spanish/Chinese.

Regarding future work, we foresee the use of this corpus in text mining applications, such as cross-language classification and clustering. In addition, the corpus could be used in parallel corpus filtering tasks, since for some of the language pairs, specially Europeans/Asian, the alignment can be noisy. For the case of Chinese/English, we can foresee the possible application of unsupervised machine translation, since the parallel patents are not usually aligned by sentences.

7. Acknowledgements

This work was supported by the Amazon AWS Cloud Credits for Research, the Google TensorFlow Research Cloud for the free credits for the usage of TPUs. The authors also acknowledge the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing HPC resources of the SDumont supercomputer, which have contributed to the research results reported within this paper.

8. Bibliographical References

Agić, Ž. and Vulić, I. (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, July. Association for Computational Linguistics.

Barnett, J., Mani, I., Martin, P., and Rich, E. (1991). Reversible machine translation: What to do when the languages don't line up. *Reversible Grammar in Natural Language Processing*.

Commons, C. (2013). Cc licenses and examples. Accessed: 2019-11-25.

Hsu, J.-A. (2014). Error classification of machine translation a corpus-based study on chinese-english patent translation. *Translation Studies Quarterly*, 18:121–136.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Kors, J. A., Clematide, S., Akhondi, S. A., Van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Lin, S.-C., Wang, J.-C., and Wang, J.-F. (2005). Translation divergence analysis and processing for mandarin-english parallel text exploitation. In *Proceedings of* *the 17th Conference on Computational Linguistics and Speech Processing*, pages 425–433.

Neves, M., Yepes, A. J., and Névéol, A. (2016). The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). 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.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Rafalovitch, A., Dale, R., et al. (2009). United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Wu, C., Xia, F., Deleger, L., and Solti, I. (2011). Statistical machine translation for biomedical text: are we there yet? In *AMIA Annual Symposium Proceedings*, volume 2011, page 1290. American Medical Informatics Association.

Zhang, S., Ling, W., and Dyer, C. (2014). Dual subtitles as parallel corpora.