Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus

Raivis Skadiņš[1], Jörg Tiedemann[2], Roberts Rozis[1] and Daiga Deksne[1]
Tilde[2], Uppsala University[2]
E-mail: raivis.skadins@tilde.lv, jorg.tiedemann@lingfil.uu.se, roberts.rozis@tilde.com, daiga.deksne@tilde.lv

Abstract

The European Union is a great source of high quality documents with translations into several languages. Parallel corpora from its publications are frequently used in various tasks, machine translation in particular. A source that has not systematically been explored yet is the EU Bookshop – an online service and archive of publications from various European institutions. The service contains a large body of publications in the 24 official of the EU. This paper describes our efforts in collecting those publications and converting them to a format that is useful for natural language processing in particular statistical machine translation. We report our procedure of crawling the website and various pre-processing steps that were necessary to clean up the data after the conversion from the original PDF files. Furthermore, we demonstrate the use of this dataset in training SMT models for English, French, German, Spanish, and Latvian.

Keywords: parallel corpus, statistical machine translation, evaluation

1. Introduction

Parallel corpora are valuable resources but not always easy to obtain and certainly not straightforward to build. The World Wide Web is a natural place to look for translated documents due to its multilingual nature. A great source of high quality documents is the collection of official publications of some legislative or administrative entity such as the Council of the European Union. A source that has not systematically been used yet is the EU Bookshop – an online service and archive of publications from various European institutions. The service is managed by the publication office of the European Union in Luxembourg and contains a large body of publications in the 24 official of the EU. This paper describes our efforts in collecting those publications and converting them to a format that is useful for natural language processing in particular statistical machine translation. We report our procedure of crawling the website and various pre-processing steps that were necessary to clean up the data after the conversion from the original PDF files. Furthermore, we demonstrate the use of this dataset in training SMT models for two use cases: (1) the case of well-resourced languages – English, German, French and Spanish, and (2) the case of an under-resourced language of the extended EU, Latvian.

2. Crawling the EU Bookshop Website

To begin with, we investigated the structure of the EU Bookshop web site[1] and found out that it is organized in such as way that it cannot easily be crawled using standard tools, but that it has an advanced search function that returns a paged list of all documents. Therefore, we created a customized web crawler that went through all pages returned by the search function and created a complete list of publications available from their web site. Each publication is connected to a title page, which contains the publication title, year of publishing, catalogue number, list of languages and other metadata. Our crawler downloaded and saved title pages of all publications and parsed them to get a list of URLs of PDF documents in all publication languages. The website is organized so that URLs of PDF documents can be composed from the publication catalogue number and the language code. Using this list, we then fetched all PDF documents using *wget*.

This procedure gave us the complete archive of the EU Bookshop with all translations available at that time together with the corresponding title page. We organized the data in such a way that each publication is stored in a separate folder named after the catalogue ID with file name extensions that specify the language of the document. Each folder also contains the appropriate title page with all its additional information provided for that publication.

3. Integration in OPUS

Our main goal is to provide a large and reasonably clean parallel corpus to the NLP community that includes all language pairs from the EU Bookshop. The main purpose of the collection is statistical MT, which basically requires large quantities of plain text documents aligned on the sentence level. Due to the enormous amounts of data in the collection, we have to rely on automatic processing without manual corrections whatsoever. However, extracting the textual content from arbitrary PDF documents is a major challenge. Even though various tools exist, there are always many cases where the conversion fails or produces unsatisfactory results. The four main steps that need to be performed are:

1. Converting PDF to plain text or XML
2. Cleaning and filtering the data

[1] http://bookshop.europa.eu/

3. Linguistic pre-processing
4. Sentence alignment

It turned out that the conversion from PDF was the toughest problem to deal with in this pipeline. PDF is a file format optimized for printing and encapsulates a complete description of the layout of a document including text, fonts, graphics and so on. It can clearly be seen that there is a large variation of documents published in the EU Bookshop combining various types of layout and visual elements created with various tools and software packages. In the end, we decided to develop our own tool, pdf2xml,[2] which combines a number of free conversion packages with several post-processing heuristics that fix text extraction errors. Our tool incorporates the PDF-rendering libraries pdfxtk[3] (Hassan, 2009a, 2009b), Apache Tika[4] (see Mattmann et al, 2011), and Poppler[5] using various configurations supported by these tools. One major issue is the detection of text unit boundaries (words and paragraphs). A typical problem is illustrated below:

```
2. Les c r i t è r e s de choix : la c o n s o m m a t i o n de c o m b u s - t ib les et l e u r moda l i t é d ' u t i l i s a t i on d 'une p a r t , la concen t r a t ion d ' a u t r e p a r t 16
```

The text above comes from a French document but the conversion failed to find proper word boundaries and instead separated most letters by additional spaces. The challenge is now to recover the text in the best possible way without heavily relying on language-specific resources. What our tool does is to use alternative conversion settings to retrieve possible textual representations; then it builds a vocabulary and a simple language model on-the-fly from those texts and runs through a baseline conversion to greedily match letter sequences that are accepted words in the vocabulary and scores them with the language model. We filter out suspiciously long words and do not merge strings that produce sequences with upper-case letters following lower-case letters. Furthermore, we include de-hyphenation heuristics using the same on-the-fly vocabulary to decide whether to keep a hyphen or not. Processing the same example from above with these techniques produces the following text:

```
2. Les critères de choix : la consommation de combustibles et leur modalité d'utilisation d'une part, la concentration d'autre part 16
```

Our conversion procedure also produces paragraph boundaries, which are helpful in subsequent processes such as sentence alignment for which we use Hunalign that mainly draws on sentence length correlations.
Another step that helps cleaning the data is language checking. Many documents contain several languages and some parts are still incorrectly converted. We decided to check each and every sentence in the corpus using our language identifier[6] (Tiedemann and Ljubešić, 2012), which is largely based on the Chrome language detection library.[7] Filtering out sentences that do not match the given language is done after sentence alignment. This is an effective way of removing further erroneous sentences from the entire data collection.
The final corpus contains over 40 languages (a few publications exist even in non-EU languages) with sentence alignments for all language pairs. Over 135,000 files are converted giving a total amount of more than 3.5 billion tokens. The collection is available in different formats (standalone XML with standoff alignment, TMX and plain Unicode UTF8 text) from the following website: http://opus.lingfil.uu.se/EUbookshop.php. The largest bitexts include over 200 million tokens in parallel. We also provide monolingual corpora for all languages to make it straightforward to train language models on the data as well (note that not all documents are translated). Table 1 below lists statistics for the ten largest languages in the collection. The upper-right part of the table shows the number of tokens for each bitext in millions (source and target language counted together) and the lower-left part shows the number of aligned sentences (also in millions). The diagonal of the table lists the total number of documents for each language.

	da	**de**	**el**	**en**	**es**	**fr**	**it**	**nl**	**pt**	**sv**
da	7081	404	367	437	347	405	424	425	323	139
de	5.01	15K	375	717	402	678	483	446	332	140
el	3.69	4.2	6486	406	373	405	395	393	350	152
en	5.09	9.62	4.09	37K	437	838	527	486	361	156
es	3.94	4.96	3.83	5.31	7716	440	406	377	351	151
fr	4.55	8.87	4.15	10.8	5.03	17K	496	451	364	159
It	4.94	6.11	4.17	6.62	4.8	5.87	9151	466	347	153
nl	5.19	5.93	4.19	6.08	4.46	5.39	5.77	7687	346	153
pt	3.66	3.99	3.56	4.25	4.04	4.1	4.08	4.04	6381	152
sv	1.79	1.84	1.54	1.96	1.78	1.83	1.86	1.93	1.81	4033

Table 1: Statistics of all bitexts for the 10 largest languages in our collection: total number of source and target language tokens in millions (upper-right), total number of aligned sentences (lower-left) and total number of documents (diagonal).

We also provide monolingual corpora for all languages to make it straightforward to train language models for the use in statistical MT. The Table 2 lists the 20 largest monolingual corpora in the collection.
Furthermore, we created some benchmarking baselines for machine translation using the popular phrase-based SMT model implemented in the open-source toolkit Moses (Koehn et al, 2007). For evaluation, we selected the news testsets from 2013 provided by the annual

[2] https://bitbucket.org/tiedemann/pdf2xml
[3] http://sourceforge.net/projects/pdfxtk/
[4] http://tika.apache.org
[5] http://poppler.freedesktop.org

[6] https://bitbucket.org/tiedemann/blacklist-classifier
[7] http://code.google.com/p/chromium-compact-language-detector/

workshop on statistical machine translation (Bojar et al, 2013). We limited our experiments to systems translating from and to English with German, French and Spanish as either input or output language. We used the standard pipeline for training translation and language models and tuned model parameters with MERT (Och, 2003) on the news testset from 2011. For efficiency reasons, we replaced GIZA++ with fast_align (Dyer et al, 2013), a drop-in replacement for word alignment that provides alignment accuracies similar to the more expensive IBM model 4 that is often used otherwise. Our language model is trained on the shuffled news dataset from 2012 using Kenlm (Heafield, 2011; Heafield et al., 2013). We did not use any additional data (like monolingual LDC corpora or other parallel resources) and did not apply advanced techniques like compound splitting, pre-ordering or lexicalized reordering.

	Documents	Sentences	Words
en	37 663	66 386 862	1 171 541 862
fr	17 261	18 473 883	445 751 294
de	15 585	18 203 612	346 449 852
it	9 151	11 073 808	265 671 485
nl	7 687	10 207 236	247 590 860
es	7 716	8 214 959	223 504 726
el	6 486	10 021 395	213 231 525
da	7 081	8 650 537	208 175 843
pt	6 381	6 975 719	184 558 611
sv	4 033	3 235 157	71 513 322
fi	4 055	3 623 953	63 068 524
pl	1 400	896 904	18 386 111
cs	1 194	848 587	16 268 410
sk	1 165	731 872	15 458 025
lv	1 165	783 181	14 915 806
hu	1 159	797 636	14 861 211
lt	1 149	826 356	14 692 008
sl	1 153	670 720	14 058 077
ro	747	541 807	13 191 342
et	1 151	720 680	12 634 912

Table 2: Statistics of monolingual corpora for the 20 largest languages in our collection.

The results are, therefore, not directly comparable with the top results obtained by other systems reported for the same test sets, which are usually heavily tuned for the specific task and the selected language pairs. For comparison, we trained additional systems with the same baseline setup but on Europarl (Koehn, 2005) data (using version 7 of that corpus). Table 3 shows the results in terms of case-sensitive BLEU scores when translating from and to English.

We can see that the EU Bookshop data makes it possible to train translation models that perform similarly well as models trained on the much cleaner Europarl data. Note that we did not perform any additional filtering to clean up the models after training. We expect that such a procedure would have a significant effect on the phrase tables considering the additional noise that appears due to the conversion from PDF. Finally, we can show that combining both data sets leads to moderate improvements showing the contribution that comes from our new data collection. Note that we did not retrain the models but simply combined the existing phrase tables using linear interpolation and tools provided by Moses. We also did not retune our SMT models for the combined models but reused weights from the Europarl baseline. Model-specific tuning would certainly improve our final system further.

System	Europarl	EU BookShop	Europarl +EU Bookshop
de-en	21.71	20.76	**22.01**
es-en	26.34	24.76	**26.52**
fr-en	26.92	26.50	**27.48**
en-de	16.13	14.69	**16.30**
en-es	25.33	23.64	**25.47**
en-fr	25.65	25.00	**26.16**

Table 3: Phrase-based SMT systems using Europarl and EU Bookshop data.

4. A Case Study in Latvian

Taking into account the big number of language pairs represented in the corpus it is not possible to do detailed evaluations for all language pairs, therefore we selected the English-Latvian language pair to conduct a deeper evaluate of its quality. We decided to do two types of evaluation: (1) manual corpus quality evaluation and (2) evaluation of its suitability for statistical MT. We selected English-Latvian because of several reasons: Latvian is a new language in EU and it is under-resourced language but still quite well represented in the corpus (ca. 0.4 mil sentences) It is a morphologically complex language and it uses several non-Latin characters which makes text extraction from PDF files more complex.

4.1 Manual Corpus Quality Evaluation

To do the manual corpus quality evaluation, we selected a random subset of 200 sentences, and asked a language expert to evaluate each sentence pair in that collection. Sentence pairs were marked as good, if sentences in both languages are more or less exact translation of each other and there are no problems with the text (encoding problems, extra spaces between characters, words merged together etc.). In all other cases the sentence pair was marked as wrong.

Initially we evaluated the version from OPUS that was prepared without any language pair specific tools using the procedures described in section 3. It contains 392,131 parallel segments, and evaluation showed that in 59.80 % of cases segments had been evaluated as good, with a confidence interval ±6.79 %. Some examples of typical errors are shown in Table 4.

English	In the European Parliament too, the European Parliament Social Economy Intergroup has been in operation since1990.In 2006 the European Parliamentcalled on the Commission "to respect thesocial economy and to present a com -munication on this cornerstone of the European social model".
Latvian	3) tās ir patstāvīgas lēmumu pieņemšanā, t.i.,var pilnīgi brīvi vēlēt un atlaist savas vadībasstruktūras un kontrolēt un organizēt to darbību;
Issue	Bad alignment
English	Results of the fight against fraud: statistical analyses and new measures taken by the Commission and the Member States .....
Latvian	Kr└pùanas apkaroùanas rezult└ti: statistikas p├t┤jumi un Komisijas un dal┤bvalstu veiktie jaunie pas└kumi .....
Issue	Broken characters
English	First indent The Commission is aware that in one national agency visited by the Court inthecontextofthisauditthe45-day deadlineforassessmentandapproval ofthefinalreportsis notrespected.
Latvian	Pirmaisievilkums.Komisijairinformētapart o,kavienā valsts aģentūrā, ko Revīzijas palāta apmeklējašīsrevīzijaskontekstā, netiek ievērots 45 dienuterminš galaziņojumu novērtēšanai un apstiprināšanai.
Issue	Lost spaces
English	• Significant reduction of internal market directives transpositiondeficit with no directives overdue by more than two years
Latvian	• B□tiski palielinjies ar iekšjo tirgu saist¥to direkt¥vu transponšanasapjoms, un kavšans direkt¥vu transponšanneprsniedz 2 gadus.
Issue	Broken characters, lost spaces

Table 4: Sample of bad segments

After the first evaluation we decided to improve the corpus quality. First, we wanted to test a different way for text extraction from PDF files and sentence alignment. We extracted the text from PDF files using Adobe PDF iFilter[8] and we lemmatized Latvian text before sentence alignment using Hunalign (Varga et al., 2005). The resulting corpus was significantly bigger, but its quality in human evaluation decreased (See EU Bookshop (v2) in Table 5).

Corpus	Size	Quality (% of good segments)
OPUS EU Bookshop (v1)	392,131	59.80 ± 6.79
EU Bookshop (v2)	561,580	44.72 ± 6.89
EU Bookshop (v3)	318,889	85.43 ± 4.88

Table 5: Results of manual corpus quality evaluation

In our third experiment we used *pdftotext* utility from Xpdf 3.03 toolkit[9] to extract the text from PDF files. After converting to plain text we compared the size of all parallel files, and if one of the files in each pair was much smaller than the other, then we discarded this pair of files from further processing assuming that sentence alignment will probably be very bad. To deal with the extra or missing spaces that frequently appear, we developed a finite state transducers for Latvian and English, which removes or inserts spaces ensuring that the output contains only correctly spelled words. For example, the transducer converts the character sequence:

```
A s  t h e  re i s  n o  re g u l a to r y  o r  fo r m a l l y  a gre e d  co n ce p t o f  s u s t a i n a b i l i t y  a p p l i c a b l e  t o  LIFE
```

to the correct text line:

```
As there is no regulatory or formally agreed concept of sustain ability applicable to LIFE
```

For sentence aligning we experimented with several tools – Microsoft's Bilingual sentence aligner (Moore, 2002), Hunalign (Varga et al., 2005), and Vanilla (Gale & Church, 1993). The alignment of Microsoft's Bilingual sentence aligner led to the most accurate results.

Finally, we filtered out parallel segments that have one of the following properties, which most likely indicate errors in the text: (1) source and target segments are completely identical, (2) a segment contains several sentences, (3) a segments is longer than 1000 characters, (4) a segment contains a word that is longer than 50 characters, (5) the source language string contains special characters of the target language or vice versa, (6) some words are written together, (7) a segment contains many short words in a row, (8) a segment contains more than 300 words, (9) the number of words in the source and the target language lines is too different, (10) bad ratio of alphanumeric characters. As some typical problems were still present, more cleaning was done. Some of the typical problems were unnecessary spaces around *fi fl* ligatures decomposed to plain character sequences, spaces around special Latvian letters scaron 'š', zcaron 'ž', imacron 'ī'. Several thousand occurrences were corrected by means of a few replace operations using regular expressions.

The resulting corpus was significantly smaller than in previous attempts, but its quality in human evaluation achieved more than 85% (EU Bookshop (v3) in Table 5).

[8] http://www.adobe.com/support/downloads/detail.jsp?ftpID=4025

[9] http://www.foolabs.com/xpdf/home.html

4.2 Training Statistical Machine Translation

Measuring quality of SMT systems has established methods, BLEU (Papineni et al., 2002) being most commonly used. This makes it possible to evaluate the downstream utility of our corpus.

In our experiment we built a phrase-based baseline English-Latvian SMT system using the DGT-TM corpus[10] (Steinberger et al., 2012) and the Moses SMT toolkit (Koehn et al., 2007), and evaluated it using a balanced evaluation set[11] containing legal texts, EU texts, news, fiction, business letters, software manuals, and popular science articles. Then we built SMT systems with each of the versions of the EU Bookshop corpus, and evaluated those, too. The results in Table 6 clearly show that the EU Bookshop corpus helps to build much better SMT systems than using just publicly available parallel EU texts. This is very important for under-resourced languages like Latvian. But the results also show that there is no direct correlation between the corpus quality and SMT quality, the correlations is rather between SMT quality and the corpus size.

SMT system	Size (= unique parallel sentences)	BLEU score
Baseline DGT-TM 2007, 2011, 2012, 2013	2.25M	13.31
Baseline + OPUS EU Bookshop (v1)	2.25M + 0.39M	19.11
Baseline + EU Bookshop (v2)	2.25M + 0.56M	**20.58**
Baseline + EU Bookshop (v3)	2.25M + 0.31M	20.33

Table 6: Results of evaluation of suitability for statistical MT

5. Conclusions

We have presented a new large parallel resource that can be applied to various NLP related tasks. Our case study demonstrates the use of our data for training statistical MT models. We have discussed various issues that we had to address when building the collection from the original PDF documents. Several steps needed to be performed to clean-up and filter the data. Data sets and tools are publically available.

6. Acknowledgements

The research leading to these results has received funding from the research project "2.6. Multilingual Machine Translation" of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and Investment and Development Agency of Latvia. The work at Uppsala University was supported by the Swedish Research Council (Vetenskapsrådet) through the project on Discourse-Oriented Machine Translation (2012- 916)

[10] releases: 2007, 2010, 2011 and 2013

[11] http://metashare.tilde.com/repository/browse/accurat-balanced-test-corpus-for-under-resourced-languages/7922fbd2a37611e3960f001dd8b71c19d96efef81e1948988b8a71b2d9d37937

7. References

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, Ph., Monz, C., Post, M., Soricut, R. and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation

Dyer, C., Chahuneau, V., Smith, N.A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL 2013*. pp. 644-648.

Gale, W.A., Church, K.W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1), pp. 75-102.

Hassan, T. (2009). GraphWrap: A system for interactive wrapping of pdf documents using graph matching techniques. In *ACM Symposium on Document Engineering*. pp. 247–248.

Hassan, T. (2009). Object-level document analysis of PDF files. In *ACM Symposium on Document Engineering*. pp. 47–55.

Heafield, K. (2011) Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 187–197.

Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 690– 696.

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

Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, In *Proceedings of the ACL 2007 Demo and Poster Sessions*. Prague, pp. 177-180.

Mattmann, C.A., Zitting, J.L. (2011). Tika in Action. Manning Publications Co. http://manning.com/mattmann/

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.

Och, F.J. (2003). Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of ACL 2003*. pp. 160–167.

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*.: ACL

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.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pp. 590-596.

Tiedemann, J., Ljubešić, N. (2012). Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*. Mumbai, India, pp. 2619–2634.