Collecting domain specific data for MT: an evaluation of the ParaCrawl pipeline

Arne Defauw, Tom Vanallemeersch, Sara Szoc, Frederic Everaert, Koen Van Winckel, Kim Scholte, Joris Brabers, Joachim Van den Bogaert
CrossLang
Kerkstraat 106
9050 Gentbrugge
Belgium
{firstname.lastname}@crosslang.com

[1] Abstract

This paper investigates the effectiveness of the ParaCrawl pipeline for collecting domain-specific training data for machine translation. We follow the different steps of the pipeline (document alignment, sentence alignment, cleaning) and add a topic-filtering component. Experiments are performed on the legal domain for the English to French and English to Irish language pairs. We evaluate the pipeline at both intrinsic (alignment quality) and extrinsic (MT performance) levels. Our results show that with this pipeline we obtain high-quality alignments and significant improvements in MT quality.

1 Introduction

In this paper, we evaluate the performance of the ParaCrawl pipeline[2] to build parallel datasets from multilingual websites related to a specific domain. The pipeline is part of the ParaCrawl[3] project mining millions of parallel sentences from the web and sharing the resulting resources online for free in all official EU languages paired with English. It starts by aligning web pages in multiple languages, applying sentence alignment for each resulting pair of web pages and a final cleaning step on the resulting sentence pairs.

The aim of this paper is to create in-domain parallel datasets by applying the existing ParaCrawl pipeline and to evaluate the resulting datasets both intrinsically (alignment quality) and extrinsically (extension of a baseline MT system with ParaCrawl results). We describe experiments on websites in the legal domain, which is sufficiently extensive to allow creating a substantial amount of domain-specific parallel data. To improve the quality of the ParaCrawl output, we add an additional topic filtering step after cleaning. We perform experiments for English-French (EN-FR) and the low resource English-Irish (EN-GA) language pairs.

2 ParaCrawl pipeline

The ParaCrawl project is co-funded by the Connecting Europe Facility and runs from 2017 to 2019. It incorporates ideas from Buck et al. (2014) and Buck and Koehn (2016a, 2016b).

Given a set of downloaded web pages (such as websites provided by the Common Crawl[4] resource, an open repository of web crawled data), the ParaCrawl pipeline performs document alignment (detection of pairs of translation-equivalent pages for two specified languages) with Malign[5], and aligns the sentences within these pairs of pages using Hunalign[6]. Finally, an additional filtering step is applied, for instance

[2] https://github.com/paracrawl, https://github.com/bitextor/bitextor

[3] https://paracrawl.eu

[4] http://commoncrawl.org

[5] Now part of https://github.com/bitextor/bitextor

[6] http://mokk.bme.hu/en/resources/hunalign

using Bicleaner[7]. The following sections describe the Malign, Hunalign and Bicleaner tools.

2.1 Malign

Considering a set of web pages in two languages[8], Malign matches web pages in the source language with translation-equivalent web pages in the target language, by detecting running text in the latter, segmenting the text and comparing the MT output of the source sentences with the target sentences. To perform this last step, an MT system is required: we trained two X>EN MT systems, one for each language pair[9].

2.2 Hunalign

Hunalign detects which sentences or groups of subsequent sentences of a document[10] in source and target language are translation-equivalents of each other. Equivalences may be 1-to-1, but also 1-to-many, many-to-1, many-to-many or null.

Alignment decisions are based on different types of information, such as sentence length and a (optionally provided, but recommended) translation dictionary. To obtain the latter, we ran GIZA++[11] on our baseline training data; from the resulting EN>X and X>EN lexical probabilities files, we generated a bilingual dictionary by multiplying the lexical probability in the EN>X direction with the probability in the X>EN direction. We retained word pairs with a lexical probability >0.1 for EN-FR, and >0.2 for EN-GA (the thresholds were obtained after manual inspection of the dictionary).

In case of alignments involving multiple sentences in one language (1-to-many or many-to-many), Hunalign will concatenate the sentences on one line in the output file. For each aligned segment, a score ranging from 0 to 1 is produced, indicating the quality of the alignment.

2.3 Bicleaner

Bicleaner detects noisy sentence pairs in a parallel corpus by estimating the likelihood of a pair of sentences being mutual translations (value near 1) or not (value near 0). Details are described in Sánchez-Cartagena et al. (2018).

Training a classifier with Bicleaner requires a clean parallel corpus (100k sentences is the recommended size) as well as source-target and target-source probabilistic bilingual dictionaries. Pre-trained classifiers for 23 language pairs[12] are already provided, including EN-FR and EN-GA.

3 Application to legal-domain data

This section describes the application of the ParaCrawl pipeline on the EN-FR and EN-GA language pairs in the legal domain. First, we describe the creation of the topic classifier and the scraping process. Then, we present and analyze the results of the latter process and of the four steps in the pipeline (document alignment, sentence alignment, cleaning and topic filtering).

3.1 Creation of fastText topic classifier

When applying the ParaCrawl pipeline for collecting domain-specific parallel data (rather than any type of bilingual material), it should be taken into account that web pages from domain-specific URLs may also contain text that is not specific to the domain of interest. Therefore, we extend the ParaCrawl pipeline with a topic filtering component. We use fastText[13] to train a model from labeled sentences by making use of sentence embeddings (Bojanowski et al. 2016, Joulin et al. 2016). We run the classifier on the output of Bicleaner and filter out sentences that are not labeled as domain-specific.

As training a fastText classifier requires labeled data, we add labels to the general and domain-specific monolingual corpora, and build a topic model for English (English being the shared source language in our experiments) to infer the topic of sentences. The data are described in Table 1. For the legal domain, we make use of the English half of the EN>FR subset of the JRC-Acquis corpus[14]. The monolingual newstest2008 dataset[15] is used as generic dataset. We retain the first 500k sentences from each corpus, deduplicate both datasets, concatenate the sentences from both datasets, and extract a held-out test set of 100k labeled sentences.

[7] https://github.com/bitextor/bicleaner

[8] Malign does not perform language classification, so the language should be specified as part of its input.

[9] Engines were trained using RNN (Recurrent Neural Network) architecture in OpenNMT (Klein et al. 2017) using the baseline training data. See section 5 for more details about the training data.

[10] Documents are split into sentences via a Moses script (see https://github.com/moses-smt).

[11] https://github.com/moses-smt/giza-pp

[12] https://github.com/bitextor/bitextor-data/releases/tag/bicleaner-v1.0

[13] https://fasttext.cc/docs/en/supervised-tutorial.html

[14] http://opus.nlpl.eu/JRC-Acquis.php

[15] http://www.statmt.org/wmt14/training-monolingual-news-crawl/news.2008.en.shuffled.gz

Domain	Data	#sentences	#retained
Legal	JRC-Acquis	814,167	470,036
Generic	newstest2008	12,954,477	497,136

Table 1: Data for topic modeling

We trained the fastText model for 25 epochs, with a learning rate of 1.0 and the wordNgrams parameter equal to 5. For other parameters we used the default settings. Our model obtains a precision and recall of 99.2% on the test set.

Based on spot-checking of the predictions on sentences from other datasets than the ones the topic model is trained with, it appears that the classifier tends to be very cautious in assigning the label “legal”. Therefore, the quality of the subset labeled as legal is very high, whereas many legal sentences are missed by the classifier. This cautiousness is also reflected by the figures for some websites: many sentence pairs resulting from scraping and aligning websites are filtered out based on the topic classifier (see Appendix A). A gold standard would be required to perform a more profound estimation of the topic model’s performance. While we did not make use of the possibility provided by fastText to assign probabilities to labels during prediction, such probabilities, in combination with a gold standard, could be used for tuning fastText towards reducing the classifier’s undershoot for the label “legal” while keeping its overkill low.

3.2 Scraping

We investigated websites in the legal domain (e.g. websites of courts) and spot-checked whether they contain information in both English as well as French and/or Irish, and whether a substantial amount of English content has a translation equivalent in one or more other languages. To make sure the scraping process would be feasible, we also took the structure of the websites into account. As for scraping tools, we use Scrapy[16], allowing to define subparts of websites to be scraped, for instance by specifying rules in a Python script to ensure only URLs with some language code in them are crawled.

A substantial amount of legal-domain content could be scraped for EN>FR, but proved to be much more difficult for the EN>GA language pair. Hence, for this language pair we decided to extend scraping to the other domains as well. While, even then, scraping resulted in a significantly smaller amount of parallel data than in the case of EN>FR, the amount of baseline data (see Table 4 in Section 5) is also modest, making the scraped data important in terms of relative size with respect to the baseline.

Table 2 shows the total number of resulting documents and sentences for each language pair. We refer to Appendix A for an overview of statistics for each scraped web-domain individually.

Pair	#doc. (EN)	#doc. (XX)	#sent. (EN)	#sent. (XX)
EN-FR	46,994	49,204	1,812,961	1,826,992
EN-GA	19,152	4,003	1,601,669	308,418

Table 2: First two columns show the number of resulting documents after scraping for each language pair and each language. Last two columns show the number of extracted sentences from these documents.

3.3 Applying the ParaCrawl pipeline

We applied the ParaCrawl pipeline (described in section 2) to the data presented in the previous section. The results are shown in Table 3. The Malign threshold was set to 0.1, and the Bicleaner threshold to 0.7 (the recommended value on the ParaCrawl project website, based on manual inspection). For Hunalign no threshold was set, so cleaning was left to Bicleaner. Again we refer to Appendix A for statistics of each web-domain individually.

Finally, after applying our topic model (see section 3.1) to the resulting corpus we obtain a domain specific corpus. We observe that a lot of sentences are filtered out by our topic model, especially for the EN-FR language pair. Looking at the results on web-domain level, this can partially be explained by the high amount of transcribed speeches scraped from the web domain www.noscommunes.ca, labeled as ‘general’ by our topic model.

	#doc. matched (Malign)	#unique aligned sent. (Hunalign)	#unique aligned sent. (Hunalign +Bicleaner)	#sent. after topic filtering (Hunalign +Bicleaner +Topic model)
EN-FR	18,808	1,472,511	786,515	79,838
EN-GA	1,575	167, 928	94,278	31,696

Table 3: Overview of the total number of documents matched with Malign, number of resulting aligned sentences after applying Hunalign (no Hunalign threshold was set),

[16] https://scrapy.org. We note that for the official release of the ParaCrawl corpus Bitextor was used for scraping (https://github.com/bitextor).

number of sentences after applying Bicleaner (Bicleaner threshold=0.7), number of Bicleaner-cleaned sentences labeled as ‘legal’ by our topic model.

4 Intrinsic evaluation

We performed an intrinsic evaluation of the aligned sentence pairs resulting from the application of the ParaCrawl pipeline to legal-domain data by comparing the pipeline’s output to a gold standard. To create the latter, we manually aligned sentences in a small subset of EN-FR and EN-GA document pairs resulting from Malign. Both automatic and manual alignment start from the same point, i.e. after document alignment and segmentation into sentences. Hence, we are not judging the document alignment component of the pipeline but merely the steps related to sentence alignment. In this section, we describe the evaluation methodology, the data used for creating the gold standard, and evaluation statistics.

4.1 Methodology

Sentence alignment involves several types of links. A typical link has a single source and a single target sentence (1-to-1 link), but there are also 1-to-many, many-to-1, many-to-many, and null links (0-to-1 or 1-to-0 links). Evaluating automatic sentence alignment takes place by comparing the output to a manually created gold standard. Manual alignment involves establishing links between one or more subsequent source sentences and one or more subsequent target sentences (Varga et al. 2005), in such a way that the links cannot be divided further into smaller links; Brown et al. (1991) refer to such sets of subsequent sentences as “beads”. The automatic sentence alignment is compared to the manual alignment based on the beads that are present in both alignments, or in just one of them. Based on this comparison, precision/recall figures can be calculated, as shown in Section 4.3. Null beads in the automatic or manual alignment are ignored during evaluation, as we do not want to bias towards this trivial type of link.

4.2 Data for Gold Standard

The gold standard was created from 13 resp. 11 document pairs for EN-FR resp. EN-GA obtained after the document alignment step described in Section 2.1.

We observed that the number of 1-to-1 beads in the Gold Standard is high, which indicates that the documents pairs are very parallel. This is not surprising, given the fact that the preceding document alignment step ignores documents that are not sufficiently parallel. We refer to Appendix B for statistics of the Gold standard.

4.3 Results

We present precision and recall scores for various thresholds of Hunalign and Bicleaner. Thresholds need to be interpreted as follows: all sentence pairs with a Hunalign probability or Bicleaner score lower than or equal to the corresponding threshold were ignored during evaluation.

To calculate recall, we take the set of gold standard beads, and the set of beads produced by the ParaCrawl sentence alignment steps for a certain threshold of Hunalign and Bicleaner. We divide the total number of shared beads by the total number of beads in the gold standard.

To calculate precision, we take the Paracrawl beads for a certain threshold of Hunalign and Bicleaner. For every bead, we look up whether it is also part of the gold standard. We divide this total number of correct predictions by the total number of predictions by the ParaCrawl pipeline for these thresholds. Precision and recall numbers for EN-FR and EN-GA are shown in Fig. 1 and Fig. 2, respectively. As we are aiming for high-quality alignments, precision is very important. Therefore, we will only consider the two rightmost columns of the matrices, which have a similar precision. These columns make clear that the Bicleaner threshold of 0.7 advised on the ParaCrawl project website is not optimal in case of our datasets: if the threshold is lowered to 0.5, the recall improves substantially.

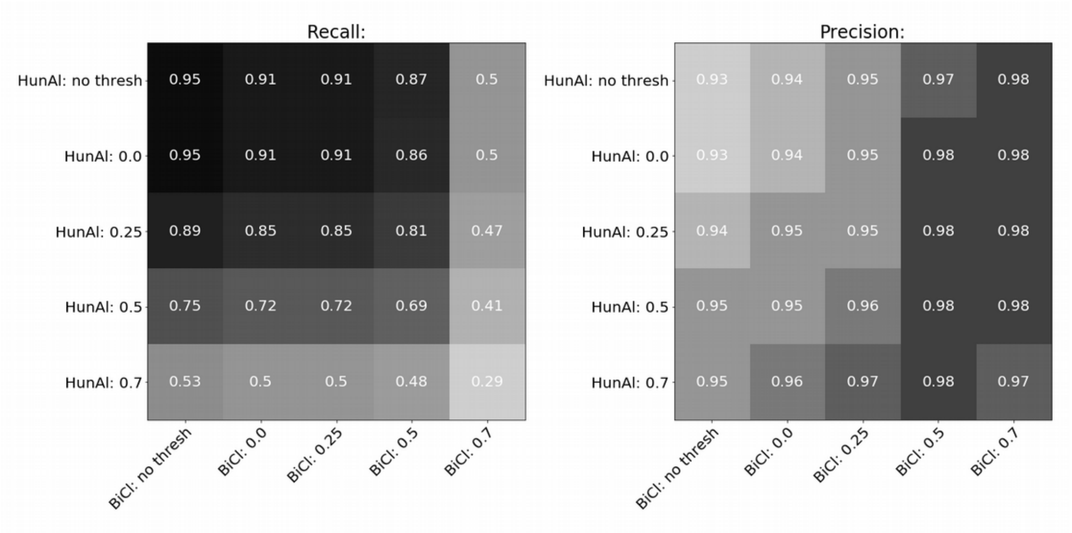

Fig. 1: Recall and precision for various Hunalign and Bicleaner thresholds (EN-FR).

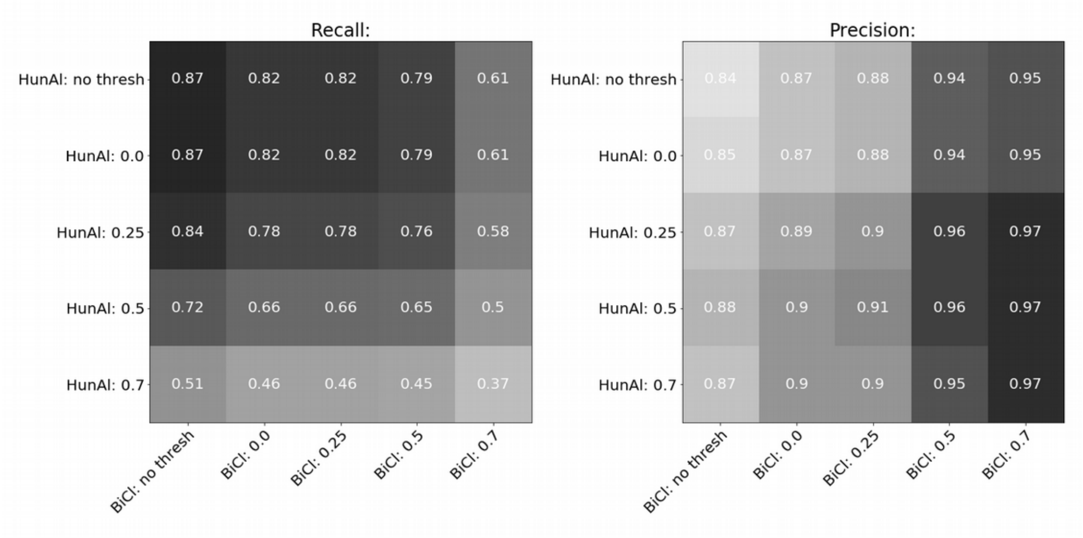

Fig. 2: Recall and precision for various Hunalign and Bicleaner thresholds (EN-GA).

5 Extrinsic evaluation

In this section, we describe the extrinsic evaluation of the parallel legal-domain data sets created with the ParaCrawl pipeline. The below sections discuss the baseline data and test sets, the training of our baseline and domain-specific MT systems, and finally the results of the baseline and domain-specific systems.

5.1 Baseline and test data

For the baseline training data, we use publicly available corpora. For EN-FR we use the *DGT*, *DCEP*, *EAC* and *ECDC* corpus, while for EN-GA also the *EUbookshop* corpus was used (see appendix C for more details). The resulting total sizes, after deduplication and removal of test sets is given in Table 4. For EN-GA we used all available parallel corpora, with the exception of legal-domain corpora (i.e. Irish legislation[17]) and of less useful corpora like *Ubuntu*.

Two types of test sets were created, in-domain and generic (see Appendix C). The in-domain test-sets were sampled from the JRC-Acquis corpus[18], the EU constitution[19] and the Irish legislation. The test set samples, consisting of unique sentence pairs, were manually verified (e.g. noisy sentences containing special layout codes or exceedingly free translations were removed). The generic test sets were sampled from the baseline data.

[17] https://www.gaois.ie/crp/en/data

[18] http://opus.nlpl.eu/JRC-Acquis.php

[19] http://opus.nlpl.eu/EUconst.php

5.2 Training of domain specific MT system

We used the generic data minus the test sets as baseline training data. In case of EN>GA, the in-domain training set has a substantial size compared to the baseline training data: 94k vs. 133k (see Table 4). The weight of the in-domain set being much lower for EN>FR (800k vs. 4M), we decided to reduce the baseline size for EN>FR to a 1M subset in order to obtain a similar weight as for EN>GA.

Type of data	EN>FR	EN>GA
Baseline training data	4,252,861	133,104
Baseline test set	3,000	3,000
Baseline sample training data	1,000,000	133,104 (sample=all)
+ in-domain training data 0.7	786,515	94,278
Total	1,786,515	227,382
+in-domain training data 0.5	1,282,978	130,807
Total	2,282,978	263,911
+in-domain training data 0.7, topic filtered	79,838	31,696
Total	1,079,838	164,800
In-domain test	3,000	3,000

Table 4: Dataset sizes (#sentence pairs).

We trained EN>FR and EN>GA Neural Machine Translation (NMT) engines with OpenNMT-tensorflow[20] using the Transformer architecture during 20 epochs and default training settings[21]. Preprocessing was done with aggressive tokenization[22], and joint subword (BPE) and vocabulary sizes set to 32k.

We concatenated the baseline training data with the in-domain data and created two domain-specific MT systems for each language pair: one based on the in-domain data produced by Bicleaner, and one on the same data, but after topic filtering (Table 4).

While we applied a threshold of 0.7 for Bicleaner, the intrinsic evaluation described in Section 4 taught us that a threshold of 0.5 provides a clearly better recall with only a slight loss in precision. Therefore, we also produced in-domain data based on Bicleaner with the lower threshold and trained a third MT-system.

5.3 Results

The translation quality of the MT models is measured by calculating BLEU scores on the two test sets. The results are listed in Table 5.

Type of data	EN>FR generic	EN>FR in-domain	EN>GA generic	EN>GA in-domain
Baseline sample training data	40.0	45.7	25.0	19.7
+In-domain 0.7	40.5	47.5	35.3	29.5
+In-domain 0.5	**41.4**	**53.1**	**37.2**	**32.8**
+In-domain 0.7, topic filtered	40.2	47.2	30.1	24.9

Table 5: Evaluation results.

These figures show that adding domain-specific training data consistently leads to improvements for both language pairs, on both generic and in-domain test sets. Nonetheless, the EN>FR systems perform clearly better on the in-domain than on the generic test set, while it is the other way around for EN>GA.

source	(vii) re-professionalisation of the military and disbanding of para-military groups,
reference	vii) reprofessionnalisation de l'armée et démantèlement des groupes paramilitaires;
baseline	vii) une reconversion de l'armée et un débarquement de groupes para-militaires,
+in-domain 0.7	vii) la réprofessionnalisation de l'armée et le démantèlement de groupes para-militaires,
+in-domain 0.7, topic-filtered	vii) la reprofessionnalisation des militaires et la dissuasion de groupes para-militaires,

Table 6: EN>FR translations of an example sentence.

When comparing the BLEU scores of the different models, it is also clear that the 0.5-0.7 range of Bicleaner adds many useful information to the parallel data, as there is a substantial increase in BLEU compared to the 0.7-1 range, especially in case of the in-domain test set (EN>FR +5.6, EN>GA +3.3). However, manual

[20] https://github.com/OpenNMT/OpenNMT-tf

[21] https://github.com/OpenNMT/OpenNMT-tf/blob/master/opennmt/models/catalog.py

[22] Standard OpenNMT tokenization but only keep sequences of the same character type, see https://github.com/OpenNMT/Tokenizer/blob/master/docs/options.md.

inspection of the output given a threshold of 0.5 teaches us that the high BLEU scores are often caused by the fact that a part of the sentence shows a strong n-gram overlap with the reference, while the remainder of the sentence is rather noisy.

As for topic filtering, the evaluation scores indicate it can be a useful step. Even though only 10% of the EN>FR domain-specific data was retained by the topic filter, the improvement in terms of BLEU (+1.5) over the baseline is almost as high as in case of adding the non-filtered data (+1.8), while much less training data is used. In case of EN>GA, the figures are different: adding the unfiltered data leads to an improvement of 9.8, whereas filtered data improves 5.2 BLEU. This difference between EN>FR and EN>GA seems to indicate that the unfiltered data for EN>FR do not add much value to the baseline data in terms of non-domain knowledge, whereas the unfiltered EN>GA data both add value in terms of non-domain and domain knowledge.

6 Conclusion

In this paper we applied the ParaCrawl-pipeline to the legal-domain: for two language pairs (EN>FR and EN>GA), we scraped a number of websites, aligned the data on document and sentence level, and added topic classification on top. We performed both intrinsic (using a gold standard) and extrinsic (by comparing a baseline MT system to domain-specific MT systems respectively) evaluations.

For the most resource-poor language pair (EN>GA), we have created a parallel resource that is substantial in size (131k) compared to publicly available data: there are 139k relevant sentence pairs on the Opus website (i.e. excluding corpora like *Ubuntu*) and 325k sentence pairs in the legal-domain. EN>GA MT systems reported on in the literature extract a much more limited amount of sentence pairs from websites or use parallel material that is not publicly available. While the EN>GA MT system Tapadóir (Dowling et al. 2015) also makes use of some websites with multilingual information, they only extracted 10k sentence pairs in total from these websites. The MT system IRIS (Arcan et al. 2016) makes use of a number of resources, among which second level textbooks (373k), which the authors received from a university but are not publicly available.

The intrinsic evaluation results show that we obtain high-quality alignments for EN-FR and EN-GA when comparing to the gold standard. We also tested different Bicleaner thresholds, which showed that 0.5 (when omitting a threshold for Hunalign) leads to a high precision and a sufficiently high recall, although both precision and recall is somewhat lower for EN-GA for all thresholds considered.

The extrinsic evaluation shows that we obtain significant improvements for both EN>FR and EN>GA when adding domain-specific data, which indicates the usefulness of the data produced by the pipeline in an MT context.

The topic filtering proved useful based on the extrinsic evaluation results. Adding only 10% of the EN>FR domain-specific data results in almost the same improvement as the one obtained when adding all data. However, this assumes a strong baseline, as indicated by the figures for EN>GA, which show a much smaller improvement when adding topic-filtered data only.

Acknowledgement

This work was performed in the framework of the SMART 2015/1091 project ("Tools and resources for CEF automated translation"), funded by the CEF Telecom programme (Connecting Europe Facility).

References

Arcan, Mihael, Caoilfhionn Lane, Eoin Ó Droighneáin, and Paul Buitelaar. 2016. IRIS: English-Irish Machine Translation System. *LREC 2016*, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 566–572.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176.

Buck, Christian, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram Counts and Language Models from the Common Crawl. *LREC 2014*, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3579–3584.

Buck, Christian and Philipp Koehn. 2016. Findings of the WMT 2016 Bilingual Document Alignment

Shared Task. *WMT 2016, Proceedings of the First Conference on Machine Translation*, Volume 2: Shared Task Papers, pages 554–563.

Buck, Christian and Philipp Koehn. 2016. Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance. *WMT 2016, Proceedings of the First Conference on Machine Translation*, Volume 2: Shared Task Papers, pages 672–678.

Dowling, Meghan, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge. 2015. Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. *LRL 2015, Proceedings of The Fourth LRL Workshop: "Language Technologies in support of Less-Resourced Languages"*, Poznan, Poland.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *ACL 2017, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:* Volume 2, Short Papers, pages 427–431.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *ACL 2017, Proceedings of ACL 2017 Conference Demo Papers,* Vancouver, Canada, pages 67-72.

Sánchez-Cartagena, Victor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. *WMT 2018, Proceedings of the Third Conference on Machine Translation*, Volume 2: Shared Task Papers, pages 995–962.

Varga, Daniel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. *RANLP 2005. Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005*, pages 590-596.

Appendices

Appendix A. Overview of corpora statistics

domain/url EN-FR	description	#doc. (EN)	#doc. (FR)	#doc. match Malign
https://e-justice.europa.eu	European e-justice portal	2,581	1,973	1,642
laws-lois.justice.gc.ca	Consolidated Acts and regulations	5,062	5,062	3,355
http://justice.gc.ca	Department of Justice	25,402	5,952	2,732
www.noscommunes.ca	House of commons	382	382	381
https://sencanada.ca	Senate	136	136	136
www.legifrance.gouv.fr	Government entity responsible for publishing legal texts online	12,110	34,378	9,270
www.oecd.org	Org. for Economic Co-operation and Developm.	1,321	1,321	1,292

Table A.1. Overview of corpora statistics on document level for each scraped web-domain (EN-FR). Last column shows the number of aligned documents using Malign (threshold=0.1).

domain/url EN-GA	description	#doc. (EN)	#doc. (GA)	#matched doc. (Malign)
https://www.education.ie	Department of Education and Skills	18,542	3,459	1,340
www.courts.ie	Courts Service	610	544	235

Table A.2. See Table A.1, but now for the EN-GA language-pair.

domain/url EN-FR	#unique aligned sent. (Hunalign)	#unique aligned sent. (Hunalign + Bicleaner)	#unique aligned sent. (Hunalign + Bicleaner +Topic model)
https://e-justice.europa.eu	50,884	26,004	16,926
laws-lois.justice.gc.ca	66,346	30,163	21,416
http://justice.gc.ca	142,458	60,841	11,785
www.noscommunes.ca	1,042,797	581,358	13,090
https://sencanada.ca	123,570	70,657	2,846
www.legifrance.gouv.fr	25,321	13,266	11,624
www.oecd.org	21,511	4,431	2,158

Table A.3: Overview of corpora statistics for each scraped web-domain (EN-FR). Second column shows the number of resulting aligned sentences after alignment with Hunalign (no Hunalign threshold was set). Third column shows results after applying Bicleaner (Bicleaner threshold=0.7). Last column shows the number of Bicleaner-cleaned sentences labeled as 'legal' by our topic model.

domain/url EN-FR	#EN tokens in unique aligned sent. (Hunalign)	#EN tokens in unique aligned sent. (Hunalign + Bicleaner)	#EN tokens in unique aligned sent. (Hunalign + Bicleaner +Topic model)
https://e-justice.europa.eu	1,376,827	690,768	496,644

laws-lois.justice.gc.ca	2,300,404	1,028,717	858,844
http://justice.gc.ca	3,571,748	1,369,891	281,943
www.noscommunes.ca	23,074,752	12,793,886	281,820
https://sencanada.ca	2.802.562	1,600,373	73,578
www.legifrance.gouv.fr	827,434	444,850	405,203
www.oecd.org	571,287	110,183	58,521

Table A.4: Overview of corpora statistics for each scraped web-domain (EN-FR). Columns show the number of EN tokens in the unique aligned sentences reported in Table A.3.

domain/url EN-GA	#unique aligned sent. (Hunalign)	#unique aligned sent. (Hunalign + Bicleaner)	#unique aligned sent. (Hunalign + Bicleaner +Topic model)
www.educationinireland.com	164,620	92,245	30,953
www.courts.ie	3,308	2,033	743

Table A.5: See Table A.3, but now for the EN-GA language pair.

domain/url EN-GA	#EN tokens in unique aligned sent. (Hunalign)	#EN tokens in unique aligned sent. (Hunalign + Bicleaner)	#EN tokens in unique aligned sent. (Hunalign + Bicleaner +Topic model)
www.educationinireland.com	4,293,616	2,615,973	961,459
www.courts.ie	78,148	50,827	21,062

Table A.6: Overview of corpora statistics for each scraped web-domain (EN-GA). Columns show the number of EN tokens in the unique aligned sentences reported in Table A.5.

Appendix B. Gold standard statistics

English Sentences	723
French sentences	716
1-to-1 beads	629
Many-to-1 beads	16
1-to-many beads	18
Many-to-many beads	1
Total number of beads used for evaluation	**664**
1-to-0 beads	35
0-to-1 beads	32
English sentences in partial links	5
French sentences in partial links	5
Total number of beads	731

Table B.1: Gold standard statistics (EN-FR). Note that partial links involve two partially equivalent sentences that are not part of a bead; they are considered as a combination of a 0-to-1 bead and a 1-to-0 bead, hence they are ignored.

English Sentences	746
Irish sentences	778
1-to-1 beads	631
Many-to-1 beads	18
1-to-many beads	19
Many-to-many beads	3
Total number of beads used for evaluation	**671**
1-to-0 beads	38
0-to-1 beads	67
English sentences in partial links	13
Irish sentences in partial links	15
Total number of beads	776

Table B.2: Gold standard statistics (EN-GA).

Appendix C. Baseline training data and test data

Corpus	EN-FR	EN-GA
DCEP[23]	3,728,978	46,418
DGT[24]	3,071,997	44,309
ECDC[25]	2,499	
EAC[26]	4,476	
Eubookshop[27]		133,363
Total (cleaned)	**4,258,861**	**139,404**

Table C.1: Overview of the training data of our baseline engines. This data was also used for training of X>EN engines necessary for document alignment.

[23] https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html

[24] http://opus.nlpl.eu/DGT.php

[25] https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory

[26] https://ec.europa.eu/jrc/en/language-technologies/eac-translation-memory

[27] http://opus.nlpl.eu/EUbookshop-v2.php

Corpus	EN-FR (full)	EN-FR (test sample)	EN-GA (full)	EN-GA (test sample)
JRC-Acquis	814,167	2000		
EU-Const	10,103	1000	10,027	1000
Acts of the Oireachtas			315,231	2000
Total	**824,270**	**3000**	**325,258**	**3000**

Table C2: Overview of the corpora used for the creation of the in-domain test sets