M1.2 – Corpora for the Machine Translation Engines

Cristina España-Bonet[1], Juliane Stiller[2] and Sophie Henning[1]
[1]Universität des Saarlandes, [2]Humboldt-Universität zu Berlin

– v4.0 –
July 2018

Abstract

This document describes the corpora used for training and evaluating the baseline MT engines used within the CLUBS project. We include as an appendix the modifications done for the final systems.

Contents

1 Introduction

The CLuBS project uses crosslingual information retrieval (CLIR) techniques to search documents in a psychology database, PubPsych[1]. These crosslingual techniques need machine translation (MT) engines in order to translate the documents. In the project, we will develop several data-based machine translation systems and study their performance as translation engines, but also their effect within the retrieval system. This document describes the first selection of corpora used for training and evaluating the systems.

Since data-based MT relies on the existence of large amount of parallel corpora, we have gathered parallel and monolingual corpora not only on psychology, but also on several other domains, some rather close in topic, such as medicine, others, such as politics, farther away. Out-of-domain corpora are easier and cheaper to get and help to improve the quality of the translators, especially when data is scarce. In our case, which is translation between De–En–Es–Fr in the domain of psychology, we can find huge amounts of data both in-domain and out-of-domain for some of the language pairs (e.g. En–De). For some other pairs only out-of domain data is available (e.g. En–Fr).

In the following sections, we describe the extraction of the monolingual and parallel corpus from the PubPsych database (Section **??**), and the monolingual and parallel out-of-domain corpus gathered from several sources (Section **??**). Section **??** presents the extraction of the evaluation corpus for the intrinsic evaluation of the machine translation. Finally, we sketch the next steps with regards to the corpus acquisition in Section **??**.

2 PubPsych Corpora

At the time of writing this report[2], the PubPsych database consists of 958.726 records. This data has been exported from the database in XML format. From all the metadata available, we use titles and abstracts for developing the translation engines. Not all the records have a title and/or an abstract and, in case they do, they are not available for all the languages in parallel. Tables **??** and **??** show the global figures of the database. English is the most populated language followed by German. Spanish and French have a similar amount of documents which is an order of magnitude smaller than English. However, although almost all the documents in French have the title also in English or Spanish, the abstract is seldom parallel. In the case of Spanish, half of the articles have the abstract both in Spanish and English. That makes French a language with almost no resources to train an in-domain MT system. For English–German and English–Spanish we can build in-domain parallel corpora, for English–French and pairs not involving English, there is not enough data.

2.1 Parallel Corpora Extraction

We select the records with titles or abstracts in more than one language to extract two parallel corpus for each language pair with a significant representation: one containing titles and one with sentences of the abstracts. In both cases, some processing of the data is needed. The code used for this processing is available in the project's GitHub repository[3].

[1] `https://www.pubpsych.eu`

[2] 1st December, 2016.

[3] `https://github.com/clubs-project/corpora-extraction-scripts`

	De	En	Es	Fr
Titles	324.005	895.982	53.065	47.707
Abstracts	250.263	513.000	34.815	33.206

Table 1: Number of records with title or abstract for the four main languages in the PubPsych database.

	En–De	En–Es	En–Fr	De–Es	De–Fr	Es–Fr	En–Es–Fr	De–En–Fr
Titles	307.37	25.680	45.324	7	50	2	2	6
Abstracts	47.218	16.934	189	0	0	105	105	0

Table 2: Multilinguality in the PubPsych database. Number of records with parallel titles or abstracts. There is no record with either the title or the abstract in all four languages.

Extraction of titles and subtitles. Records sometimes not only have a title but also a subtitle in a separate field. For building the parallel corpus one has to align these fields between the languages L1 and L2, because the title information in one language may be split in separate title and subtitle in another language. In general, one can consider four different cases:

(i) Only the title is available for L1 and L2

(ii) L1 has title and subtitle, L2 only has title

(iii) L1 only has title, L2 has title and subtitle

(iv) Both L1 and L2 have title and subtitle

In order to extract the appropriate pairs, we implement a heuristic based on the expected ratio of the length of sentences (length factor) between the two languages. We assume that the length factor follows a normal distribution and estimate its mean and standard deviation in out-of-domain parallel corpora[4]. Then, for a given parallel record, we generate all possible combinations between the available titles and subtitles and extract the pair with the closest length factor to the expected mean of the language pair.

We notice that all the cases where the title information is split into title and subtitle in PubPsych are from the Psyndex source database. The division can either be in the title of the English side or in the German one. Besides, there are no records where both L1 and L2 have title and subtitle.

Extraction of parallel sentences in abstracts. Parallel abstracts have to be aligned at sentence level in order to build the MT corpus. To do this, sentences are first split with an in-house sentence splitter and then passed to an aligner based on lengths and positions within the document. We use an available Python implementation of the Gale-Church algorithm[5] [**?**] for this. After the alignment, only sentences with more than 5 words are used for the parallel corpus of abstracts; the restriction does not apply to titles. In both cases, sentences completely uppercased are converted into lowercase strings.

2.2 Monolingual Corpora Extraction

For building the monolingual corpora, we consider all the records in the database with either a title and/or an abstract and extract the text in English, French, German and

[4]See corpora in Section **??**.

[5]`https://github.com/alvations/NTU-MC/blob/master/ntumc/toolkit/gale_church.py`

Abstracts

	En–De			En–Es			En–Fr		
	Sent.	EnTok.	DeTok.	Sent.	EnTok.	EsTok.	Sent.	EnTok.	FrTok.
Train	241.749	6.584.364	6.135.612	88.848	2.640.441	2.909.559	0	0	0
Dev.	1.500	39.968	37.557	1.500	45.611	50.831	0	0	0
Test	2.162	60.219	55.610	2.486	74.382	81.575	823	25.884	29.226

Titles

	En–De			En–Es			En–Fr		
	Sent.	EnTok.	DeTok.	Sent.	EnTok.	EsTok.	Sent.	EnTok.	FrTok.
Train	306.640	3.480.727	3.059.048	25.105	293.164	340.203	45.137	463.610	567.618
Dev.	0	0	0	0	0	0	0	0	0
Test	737	9.691	8.202	575	6.935	8.002	187	2.589	3.012

Table 3: Size of the parallel corpora obtained from the PubPsych database and used for training, development and testing the MT systems.

Spanish. For abstracts, text snippets are split into sentences. For titles, both the title and the subtitle fields are extracted. We do not need to deal with alignment issues in this case and all the sentences can be gathered.

2.3 Training, Development and Test Sets

From the full corpus, we first put aside 800 records that will be translated manually to build a high-quality test set. These records are selected from the subset of documents with an English abstract only (i.e. without any translation) and then sampled randomly according to the weight of its data source (e.g. PSYNDEX, ISOC-Psicología, etc.). The remaining MT corpus consists of 957.926 documents.

We create a partition with 1500 documents on the MT corpus to build the development and test sets. For this partition, we give preference to records with multilingual abstracts and again sample the MT corpus randomly according to the weight of its data source. The division between development and test is done differently for every language pair with the purpose of having a similar amount of parallel sentences. We use 1500 parallel sentences from the abstracts for development and the rest for testing. All the titles are used to build a test set and no development set is considered as a first approximation. In the case of French, we can only build one set due to the small amount of parallel data. The remaining 951.926 documents with title and/or abstract are used for training.

Table **??** shows the sizes for the in-domain parallel corpora obtained for each language pair after the division in training, development and test. Table **??** shows the sizes of the in-domain monolingual corpora for the four languages involved; in this case, all data is used for training.

2.4 Pre-Processing

In order to use the previous corpora in natural language applications, text must be pre-processed. We follow a three-step pre-processing standard in MT here:

1. Text normalisation
2. Text tokenisation
3. Truecasing

Abstracts & Titles

	De		En		Es		Fr	
	Sent.	Tok.	Sent.	Tok.	Sent.	Tok.	Sent.	Tok.
Abst.	1.618.845	36.093.889	3.533.855	92.211.688	168.423	5.393.989	150.537	4.571.979
Titles	396.152	3.343.386	896.876	10.741.170	52.399	631.757	47.529	595.249

Table 4: Size of the monolingual corpora obtained from the PubPsych database and used for training the MT systems.

For text normalisation and text tokenisation we use scripts from the Moses toolkit [**?**]. The normaliser, however, has been adapted to deal with some peculiarities of the PubPsych corpus. For truecasing, we have trained statistical models using Wikipedia and Europarl V7 monolingual corpora (Section **??**).

3 Out-of-Domain Corpora

The data described in the previous section is not enough to train high-quality translators, but it is important to enrich larger corpora with the language of psychology itself. In the following, we describe several open parallel corpus that will be used within the project to train general and specialised translators.

3.1 Medicine, Biological and Health domains

First, we describe two corpora that can be considered to be rather close in domain to psychology: the EMEA [**?**] and the Scielo corpus.

EMEA Corpus. This is a parallel corpus made out of PDF documents from the European MEdicines Agency[6]. Files were automatically converted from PDF to plain text, sentence-aligned and made publicly available within the Opus Corpus [**?**]. The corpus is available in 22 languages including English, French, German and Spanish. At the moment, we only consider the language pairs involving English to be compatible with the other corpora, but this corpus could be used later in the project for all the language pairs.

Scielo Corpus. This is a parallel corpus made out of documents retrieved from the SCIentific Electronic Library Online[7]. The documents belong to the biological and health domains and can be composed of either a title, an abstract or both of them. The corpus was prepared by the organisers of the Biomedical Translation Task in the First Conference on Machine Translation[8] (WMT16), who aligned the documents at sentence level with the GMA tool[9]. In this case, the corpus is available for the English–Spanish and English–French language pairs.

[6] `http://www.emea.europa.eu`
[7] `http://www.scielo.org`
[8] `http://www.statmt.org/wmt16/biomedical-translation-task.html`
[9] `http://nlp.cs.nyu.edu/GMA`

	En–De			En–Es			En–Fr		
	Sent.	EnTok.	DeTok.	Sent.	EnTok.	EsTok.	Sent.	EnTok.	FrTok.
UN	162.981	6.098.083	5.617.876	11.196.913	320.064.682	366.072.923	12.886.831	361.877.676	421.687.471
EP	1.920.209	53.091.548	50.548.739	1.965.734	54.505.707	57.047.216	2.007.723	55.730.752	61.888.789
ComCrawl	2.399.123	58.864.439	54.570.779	1.845.286	46.855.705	49.557.537	3.244.152	81.084.856	91.281.890
subTOTAL	*4.482.313*	*118.054.070*	*110.737.394*	*15.007.933*	*421.426.094*	*472.677.676*	*18.138.706*	*498.693.284*	*574.858.150*
EMEA	1.108.752	14.477.119	13.197.725	1.098.333	14.334.648	15.975.506	1.092.568	14.317.365	17.046.979
ScieloBio	–	–	–	117.862	3.252.183	3.382.511	–	–	–
ScieloHealth	–	–	–	558.714	14.382.853	15.031.533	9.129	244.486	308.055
subTOTAL	*1.108.752*	*14.477.119*	*13.197.725*	*1.774.909*	*31.969.684*	*34.389.550*	*1.101.697*	*14.561.851*	*17.355.034*
PubPsych	241.749	6.584.364	6.135.612	88.848	2.640.441	2.909.559	–	–	–
TOTAL	*5.832.814*	*139.115.553*	*130.070.731*	*16.871.690*	*456.036.219*	*509.976.785*	*19.240.403*	*513.255.135*	*592.213.184*

Table 5: Size of the parallel corpora obtained from the different sources described in Section **??**: United Nations (UN), Europarl V7 (EP), Common Crawl (ComCrawl), EMEA and Scielo. Figures of the PubPsych corpus are shown for comparison.

	De		En		Es		Fr	
	Sent.	Tok.	Sent.	Tok.	Sent.	Tok.	Sent.	Tok.
Wikipedia	39.036.439	675.868.710	92.284.575	1.920.645.814	20.085.435	465.828.442	26.603.296	553.201.962
General	4.482.313	110.737.394	18.138.706	498.693.284	15.007.933	472.677.676	19.240.403	592.213.184
Medicine	1.108.752	13.197.725	1.774.909	31.969.684	1.774.909	34.389.550	1.101.697	17.355.034
PubPsych	1.618.845	36.093.889	3.533.855	92.211.688	168.423	5.393.989	150.537	4.571.979

Table 6: Size of the monolingual corpora obtained from the different sources described in Section **??**. *General* includes the UN, EP and ComCrawl corpora and *Medicine* includes the EMEA and Scielo ones.

	En–De			En–Es			En–Fr		
	Sent.	EnTok.	DeTok.	Sent.	EnTok.	EsTok.	Sent.	EnTok.	FrTok.
news-test2012	3.003	72.988	72.603	3.003	72.988	78.887	3.003	72.988	81.797
news-test2013	3.000	64.809	63.411	3.000	64.809	70.540	3.000	64.809	73.658
EMEAdev	2.000	38.658	37.945	2.000	36.676	39.959	2.000	34.554	41.026
EMEAtest	2.000	36.864	35.773	2.000	34.359	38.615	2.000	33.316	39.674
PubPsychDev	1.500	39.968	37.557	1.500	45.611	50.831	–	–	–
PubPsychTest	2.162	60.219	55.610	2.486	74.382	81.575	823	25.884	29.226

Table 7: Size of the out-of-domain test sets to be used in the project. Figures of the automatic PubPsych test sets are shown for comparison.

3.2 General Corpora: Web and Politics

Second, we use corpora from very distant domains to gather general phrases. These corpora include texts on politics (Europarl Corpus [**?**] and United Nations Corpus [**?**]) and crawls from the web (Common Crawl corpus and Wikipedia articles). Similarly to EMEA, we only consider the language pairs involving English in the Europarl (EP) and United Nations (UN) corpora, but they could be used later in the project for all the language pairs. The Common Crawl parallel corpus is made publicly available by the annual Shared Task on Machine Translation (WMT), currently Conference on Machine Translation. The Wikipedia corpus is a monolingual in-house corpus with dumps downloaded from Wikipedia[10] in January 2015 and pre-processed with JWPL[11] [**?**].

Tables **??** and **??** show a summary of the amount of data available for every language pair. English–French is the pair with most parallel data, more than 19 million parallel sentences, but all of them belong to out-of-domain data. EMEA and Scielo are the closest sources in this case to obtain in-domain vocabulary together with the addition of the thesauri used in psychology (see Deliverable 3.1). On the other side lies the English–German language pair. Here we have less than 6 million parallel sentences but almost 250 thousand belonging to PubPsych. Finally, there are almost 17 million parallel sentences for the English–Spanish pair, with less than 100 thousand belonging to psychology.

For the second version of the corpus, we include data for the language pairs not involving English (See Apendix **??**) which is only available for the out-of-domain setting. We use Europarl [**?**], United Nations Corpus [**?**], JR-ACQUIS [**?**] and News Commentary [**?**] for that purpose. For somehow alleviating the scarcity of data in German, we also include the MODEL Rapid corpus [**?**] for all the language pairs involving German. Finally, we include the first 2 million parallel sentences from the EUbookshop [**?**] for German–Spanish and German–French. The latter two sources are supposed to be more noisy than the previous ones and for this reason we only include them for the pairs with less data.

3.3 Development and Test Sets

For development and testing we consider four sets, two corresponding to what we have called general domain and two more from the medical domain. For the general domain we use *news-test2012* and *news-test2013*. These test sets are distributed by the organisation of the WMT workshops. The years selected for our experiments are the last ones to include English, French, German and Spanish simultaneously, so that the sets are aligned among the four languages. For the medical domain we use a subset of the EMEA corpus. In this case, the sets are only aligned inside a language pair.

[10] `https://dumps.wikimedia.org`

[11] `https://code.google.com/p/jwpl`

Data source	# of records	# of recs with 1 EN abstract	# for sample
ERIC	107.532	104.648	187 (23%)
PSYNDEX	331.469	56.202	100 (13%)
PASCAL	206.670	132.439	236 (30%)
NARCIS	29.847	13.734	24 (3%)
ISOC	50.275	705	1 (0%)
PSYCHOPEN	1.059	976	2 (0%)
PSYCHDATA	53	1	0 (0%)
NORART	11.443	0	0 (0%)
TOTAL	958.726	448.732	800 (100%)

Table 8: Number of extracted documents for the evaluation corpus based on source weighting.

Table **??** shows the figures for these dev/test sets and the comparison with the test sets gathered automatically from the PubPsych data. The upcoming human translated test set will allow the evaluation of all the language pairs, also for in-domain translations.

3.4 Pre-Processing

All the corpora introduced in this section has been pre-processed with the same pipeline and tools outlined in Section **??**.

4 Corpora for Evaluation

To intrinsically evaluate the machine translation with scores like BLEU [**?**], a corpus of 800 (metadata)records from PubPsych aligned in the four languages English, Spanish, German and French is provided. For that, 800 records with only one English abstract were extracted from the full corpus. They will be translated into the respective other three languages in the near future.

The corpus for evaluation was extracted from the full data file provided on Nov. 19, 2016, with a total of 958.726 records. From these, we extracted the ones with only one English abstract and no parallel translations resulting in a corpus of 448.764 records. From these records, we sampled 800 records randomly based on the weighting of the data source. Table **??** lists the properties per source database and the percentage and corresponding number of records in the evaluation corpus.

Table **??** shows stats on the extracted corpus grouped by the publication language.

5 Conclusions

We have gathered a first collection of corpora to build the baseline MT systems for the project. The performance of the baseline systems will help us to detect new necessities on corpora in order to improve the systems. Possible weak points of the current selection include:

(i) SMT engines built using pivot methodologies do not perform well enough. In this case, we should consider German–French, German–Spanish, French–Spanish corpora when they are available

pub_lang	records	subtitle_en	title_de	title_en	title_es	title_fr
chi	1	0	0	1	0	0
dut	1	0	0	1	0	0
en	24	0	0	24	0	0
eng	711	3	87	711	1	0
fre	7	0	0	6	0	7
ger	11	0	11	11	0	0
ita	1	0	0	1	0	0
jpn	2	0	0	2	0	0
por	1	0	0	1	0	0
spa	4	0	0	4	4	0

Table 9: Publication language of original documents in the evaluation corpus.

(ii) MT engines show a very low quality when translating titles. In this case, we should train specific systems for titles and therefore prepare additional development/test sets

(iii) We observe some untranslated words due to truecasing. In this case, a lowercased version of the corpora would be used.

And already detected necessities:

(i) PubPsych human translated test sets fully aligned to test in-domain translation in all the language pairs

(ii) Multilingual thesauri terms to add to current parallel corpora

A Additional Corpora

For the last version of translator we have gathered additional corpora in German–English, German–French, German–Spanish and French–Spanish in order to improve the weak points of the previous corpus reported in the conclusions of the document. These data are only out-of-domain as PubPshyc does not have aligned documents for these pairs.

We have also improved the pre-processing of the data after some issues detected in the first corpus. The current pre-processing includes:

1. Removal of sentences where more tha 50% of the text is in a non-latin alphabet (this removes equations and sentences in other languages such as Greek or Chinese)
2. HTML entities cleaning (for instance, correcting double conversions such as apos-aposquote; instead of &aposquote;)
3. Text normalisation
4. Text tokenisation
5. Truecasing
6. Duplicate removal

The full pre-processing pipeline is available in the git repository of the project[12]. Major differences come for the EMEA data, a corpus with lots of duplicate parallel sentences and equations. In this case, we finally get a clean subcorpus with only a quarter of the orginal sentences.

[12] `https://github.com/clubs-project/corpora-extraction-scripts/tree/master/cleaning`

A.1 Parallel Corpora Figures

	En–De			En–Es			En–Fr		
	Sent.	EnTok.	DeTok.	Sent.	EnTok.	EsTok.	Sent.	EnTok.	FrTok.
UN	114.085	5.001.920	4.555.627	8.229.875	256.932.146	294.068.300	9.091.417	282.087.198	329.282.596
EP	1.862.433	52.674.125	50.126.351	1.910.098	54.065.460	56.618.936	1.956.314	55.312.925	60.886.010
ComCrawl	2.394.117	58.649.429	54.440.232	1.840.956	46.659.115	49.438.475	3.231.507	80.589.534	89.562.986
NC	221.211	5.502.983	5.608.068	236.969	6.019.121	6.881.068	207.894	5.237.637	6.240.595
JRC	473.128	15.629.431	14.050.433	–	–	–	–	–	–
Rapid	1.010.295	22.707.408	22.151.469	–	–	–	–	–	–
subTOTAL	*6.075.269*	*160.165.296*	*150.932.180*	*12.217.898*	*363.675.842*	*407.006.779*	*14.487.132*	*423.227.294*	*485.972.187*
EMEA	237.607	4.111.481	3.779.495	240.888	4.229.668	4.697.635	246.315	4.258.733	5.036.208
ScieloBio	–	–	–	116.999	3.246.340	3.376.848	–	–	–
ScieloHealth	–	–	–	532.287	14.016.862	14.649.695	8.990	242.750	304.494
subTOTAL	*237.607*	*4.111.481*	*3.779.495*	*890.174*	*2.149.2870*	*22.724.178*	*255.305*	*4.501.483*	*5.340.702*
PubPsych	542.690	10.023.347	9.161.397	112.148	2.907.160	3.222.127	44.505	457.745	560.716
TOTAL	*6.753.879*	*148.073.799*	*137.456.811*	*13.220.220*	*388.075.872*	*432.953.084*	*14.786.942*	*428.186.522*	*491.873.605*

	De–Es			De–Fr			Es–Fr		
	Sent.	DeTok.	EsTok.	Sent.	DeTok.	FrTok.	Sent.	EsTok.	FrTok.
UN	120.334	4.773.923	6.115.036	120.529	4.736.002	6.111.522	8.488.449	307.870.137	314.689.536
EP	1.842.324	49.824.584	54.749.655	1.901.615	51.051.092	58.973.416	1.937.659	57.569.936	60.481.886
JRC	963.545	28.133.439	33.757.484	951.566	27.759.441	33.481.936	962.918	33.533.403	33.674.993
NC	209.534	5.453.601	6.119.785	185.039	4.763.166	5.581.034	194.943	5.763.484	6.008.991
Rapid	526.605	11.499.817	13.484.752	975.509	21.045.994	25.922.810	–	–	–
subTOTAL	*3.662.342*	*99.685.364*	*114.226.712*	*4.134.258*	*109.355.695*	*130.070.718*	*11.583.969*	*404.736.960*	*414.855.406*
EMEA	258.428	4.138.271	4.988.972	265.202	4.171.805	5.354.591	265.230	5.117.314	5.472.197
TOTAL	*3.920.770*	*103.823.635*	*119.215.684*	*4.399.460*	*113.527.500*	*135.425.309*	*11.849.199*	*409.854.274*	*420.327.603*

Table 10: Size of the parallel corpora after the new cleaning pipeline obtained from the different sources described in Section **??** with (Multi-) United Nations (UN), Europarl V7 (EP), Common Crawl (ComCrawl), JRC-Acquis (JRC), MODEL Rapid (Rapid), EMEA and Scielo. Figures of the PubPsych corpus are shown for comparison.