The Scielo Corpus:
a Parallel Corpus of Scientific Publications for Biomedicine

Mariana Neves[*], **Antonio Jimeno Yepes**[†], **Aurélie Névéol**[‡]
[*]Hasso-Plattner Institute
August-Bebel-Str. 88 14482 Potsdam - Germany
mariana.neves@hpi.de

[†]IBM Research Australia, Carlton 3053 VIC - Australia,
University of Melbourne, Parkville 3010 VIC - Australia
antonio.jimeno@au1.ibm.com

[‡]LIMSI, CNRS, UPR 3251 Université Paris-Saclay F-91405 Orsay - France
aurelie.neveol@limsi.fr

Abstract

The biomedical scientific literature is a rich source of information not only in the English language, for which it is more abundant, but also in other languages, such as Portuguese, Spanish and French. We present the first freely available parallel corpus of scientific publications for the biomedical domain. Documents from the ”Biological Sciences” and ”Health Sciences” categories were retrieved from the Scielo database and parallel titles and abstracts are available for the following language pairs: Portuguese/English (about 86,000 documents in total), Spanish/English (about 95,000 documents) and French/English (about 2,000 documents). Additionally, monolingual data was also collected for all four languages. Sentences in the parallel corpus were automatically aligned and a manual analysis of 200 documents by native experts found that a minimum of 79% of sentences were correctly aligned in all language pairs. We demonstrate the utility of the corpus by running baseline machine translation experiments. We show that for all language pairs, a statistical machine translation system trained on the parallel corpora achieves performance that rivals or exceeds the state of the art in the biomedical domain. Furthermore, the corpora are currently being used in the biomedical task in the First Conference on Machine Translation (WMT'16).

Keywords: parallel corpus, machine translation, biomedicine

1. Introduction

Researchers make available new findings and knowledge in biology and medicine through the publication of research articles. However, the volume of the biomedical literature grows at a large rate, which poses a challenge to keep up with new discoveries. Even though a large volume of papers are published in English, there are also many publications written in languages other than English.

Access to the biomedical literature is available on-line via systems such as PubMed®[1], that allow researchers to browse and search for publications of their interest. Even though English is the *de facto* official language in the scientific community, many researchers have limited proficiency in English and feel more comfortable with reading scientific prose written in their native language. The same goes for patients who might find it difficult to understand medical records when they are written in a language other than their native language. Furthermore, articles published in local research journals are accessible only for researchers fluent in the original language of the article. This is typically the case for articles available in databases such as Scielo[2], which has a focus on Latin American publications.

Machine translation (MT) can provide a solution to increase the access to the biomedical literature (Pecina et al., 2014) and to health information in general (Kirchhoff et al., 2011). For instance, it was successfully used as the basis for query translation in cross-language information retrieval (Pecina et al., 2014). Although there has been much work in this field (Bojar et al., 2014), the automatic translation of scientific publications has not received much attention of the community, in part because of the difficulty of getting parallel collections of documents.

PubMed is the largest database for scientific publications in biomedicine, and has been extensively used in many biomedical natural language processing applications. However, only titles are available in more than one language in PubMed (Wu et al., 2011). Previous work (Jimeno Yepes et al., 2013) relied on titles extracted from PubMed and abstract texts extracted from various journals' web sites for experiments on MT, but licenses issues limited the distribution of the corpora used in the work.

Herein, we describe the construction of the first freely available parallel collection of scientific publications for the biomedical domain. The documents were derived from Scielo[3], a database of open access scientific publications with a focus on developing and emerging countries, and especially on Latin America. Scielo currently includes publications in a variety of domains, such as agriculture, engineering, biological and health sciences. It includes abstracts and full texts for publications, mainly in Portuguese and Spanish, but also in English, French and German.

The intended purpose of this parallel corpus is to train and evaluate MT systems. To this end, we created parallel corpora for three pairs of languages: Spanish-English

[1] http://www.ncbi.nlm.nih.gov/pubmed

[2] http://www.scielo.org/

[3] http://www.scielo.org/

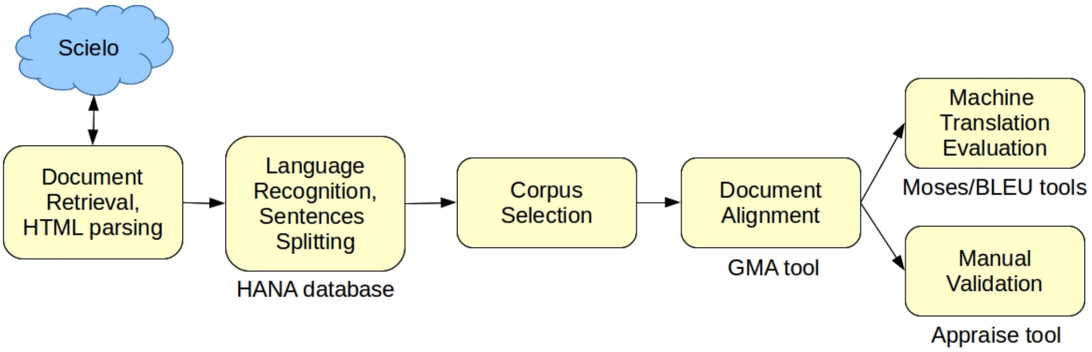


Figure 1: Work-flow of the construction of the parallel collection of biomedical publications.

(ES/EN), French-English (FR/EN) and Portuguese-English (PT/EN). These collections are used as training data for the biomedical shared task[4] in the First Conference on Machine Translation (WMT16). In this paper, we describe the set of documents which was collected, though we have currently only released the training dataset, while the test dataset is being kept for future release during the shared task. Examples of the sentences for all language pairs, i.e., ES/EN, FR/EN and PT/EN, are shown below:

La especie más frecuente aislada de pacientes de ambas regiones fue L. paracasei ssp paracasei 1.
Lactobacillus paracasei ssp paracasei 1 was the most frequently isolated species in both regions.

Le seul traitement validé pour soigner cet état est l'immunothérapie passive avec des sérums antivenimeux d'origine animale sûrs et efficaces.
The only validated treatment for this condition is passive immunotherapy with safe and effective animal-derived antivenoms.

Avaliação da força muscular periférica de pacientes submetidos à cirurgia cardíaca eletiva: estudo longitudinal.
Evaluation of peripheral muscle strength of patients undergoing elective cardiac surgery: a longitudinal study.

2. Related Work

The development of parallel corpora as translation memories and use for training MT systems has been an active area of research. Previous work has addressed various types of documents, domains and language pairs. Popular parallel corpora in specialized domains include the News Commentary corpus[5], composed of news related documents and commonly used in the WMT challenges, the EuroParl corpus (Koehn, 2005), derived from the European Parliament proceedings, and the Acquis corpus (Steinberger et al., 2006), which contains legal documents for more than 20 European languages.

The OPUS corpus (Tiedemann, 2012) contains some subdomain data covering the biomedical domain with EMEA (European Medicines Agency) documents, which were later used, along with MEDLINE® titles, to produce parallel annotated data following the CLEF-ER 2013 challenge (Kors et al., 2015) and for named-entity recognition and normalization for French in the QUAERO corpus (Névéol et al., 2014). Previous attempts to train statistical MT systems for biomedicine include the pioneer work using a corpus of MEDLINE titles (Wu et al., 2011), use of the Cochrane Systematic Reviews for the FR/EN language pair (Névéol et al., 2013), use of publication titles and abstracts for both ES/EN and FR/EN language pairs (Jimeno Yepes et al., 2013) and building an application to produce public health information (Kirchhoff et al., 2011). In summary, currently, there are limited resources for training and evaluating MT systems for the biomedical domain as previous work is limited to either document titles or one particular language pair, e.g., FR/EN.

3. Methods

In this section we present our methodology for constructing parallel corpus of scientific publication for biomedicine derived from the Scielo database. The work-flow is illustrated in Figure 1 and each phase is described in details below.

3.1. Document retrieval and HTML parsing

We developed a crawler for the Scielo web site and retrieved articles periodically from Scielo. Our crawling has its starting point in the pages that list all journals from the "Biological Sciences" and "Health Sciences" subjects. These categories are used to compose the two datasets, with the corresponding names, of our corpus. Despite being distinct categories in Scielo, these are overlapping categories, as there are many journals that belong to both of them.
From the list of journals, it is possible to retrieve a list of all issues of a particular journal, which is available in the regional web sites of Scielo in distinct countries, such as Brazil, Chile or Colombia. The HTML page of the journal's list of issues was further parsed to retrieve the page which contains the list of articles of a given issue. Finally, we downloaded the page of a particular article and parsed the HTML code in order to extract the title and the abstract of

[4] http://www.statmt.org/wmt16/biomedical-translation-task.html
[5] http://www.statmt.org/wmt16/translation-task.html

each publication. Titles and abstracts were subsequently stored and indexed in the SAP HANA database[6].

All translations of the abstracts in Scielo are the original texts provided by the authors of the publications, who are presumably not professional translators, and who may not have native proficiency in both languages. All publications are available under either the Creative Commons Attribution-Noncommercial 3.0 Unported (cc-by-nc) or Attribution 3.0 Unported (cc-by) licenses, which makes all documents suitable for redistribution and research purposes.

3.2. Language recognition and Sentence Splitting

We used the HANA database to perform language recognition in the texts and their segmentation into sentences. Although the language of the publication is usually identified in the Scielo URL, we noticed that there are many situations in which the abstract is in one language and the title in another, making the language recognition step necessary. For instance, the document S0874-48902010000300006[7] contains the abstracts available in French, Spanish and English, four different HTML pages, but the title is always in Portuguese in all of them.

We decided to use the HANA database tool for sentence splitting after finding that it compared favorably to the OpenNLP library[8] on a sample of documents. Further, as stated above, HANA could also be used for language recognition and provides support of various languages, including the ones we focus on in this work.

Corpus Selection

For both "Biological Sciences" and "Health Sciences" categories, we retrieved from the database pairs of titles and abstracts available in both English and one of the other three languages we consider, i.e., French, Portuguese or Spanish. These constitute our whole collection of parallel documents, which was subsequently split in training and test datasets.

We built the test datasets by retrieving 1,000 complete documents, i.e., containing both title and abstract, from our database, 500 for each of the translation directions, e.g., English to Spanish and Spanish to English. The only exception is the FR/EN language pair for Biological Sciences for which very few parallel documents are available. When selecting the documents of the test dataset, we checked that none of them is included in the training or test dataset of the other language pairs, categories (Biological and Health) and language directions. Once the test datasets were selected, the rest of the parallel documents, i.e., complete documents, abstracts or titles, constitutes the training datasets.

Scielo contains many entries only available in one of the languages or in languages other than English, e.g., in both Portuguese and Spanish, given that the focus of the database is in the Latin American journals. These documents constitute our monolingual corpus, given that in-domain monolingual corpora are also a valuable resource for training and evaluation of language models, one of the components of statistical MT systems (Koehn, 2010).

3.3. Document alignment

We automatically aligned sentences from titles and abstracts for the language pairs using the Geometric Mapping and Alignment (GMA) tool[9]. No language-specific resources were provided while using the GMA tool, such as bilingual dictionaries. As discussed above, many articles in Scielo do not have both title and abstract available for a particular language, but just one of them. For this reason, we decided to align titles and abstracts separately.

3.4. Manual validation

We manually checked the automatic alignment generated by the GMA tool to ensure the quality of the corpora. Statistical MT tools need to rely not only on parallel collections of documents, but also on parallel collections of aligned sentences. We randomly selected 100 publications (titles and abstracts) for each category, i.e., Biological Sciences and Health Sciences, and for each of the three pairs of languages, i.e., PT/EN, ES/EN and FR/EN. The sentences were converted into the XML format of the Appraise tool[10] and loaded into the tool. Each of the authors validated the documents pair for his or her native language, i.e., MN for Portuguese, AJY for Spanish and AN for French. The sentences were validated using the "Quality Checking" task (cf. Figure 2) available in Appraise to check the degree of alignment between each pair of sentences. We defined five categories to classify the alignment: (a) "OK", when the alignment is correct and both sentences contains the same information; (b) "Source>Target", when the alignment is correct but the source contains more information than the target; (c) "Target>Source", when the alignment is correct but the target contains more information than the source; (d) "Overlap", when there is an overlap in the information content of both sentences but they cannot be considered aligned; (e) "No alignment", when the sentences are unrelated and there is no alignment.

3.5. Machine Translation Evaluation

We trained a statistical MT system on the parallel corpora to demonstrate the capabilities of the proposed corpus, For this purpose, we used Moses[11](Koehn et al., 2007) as the statistical MT tool. The BLEU score (Papineni et al., 2002) has been used as the translation evaluation measure.

[6] http://hana.sap.com

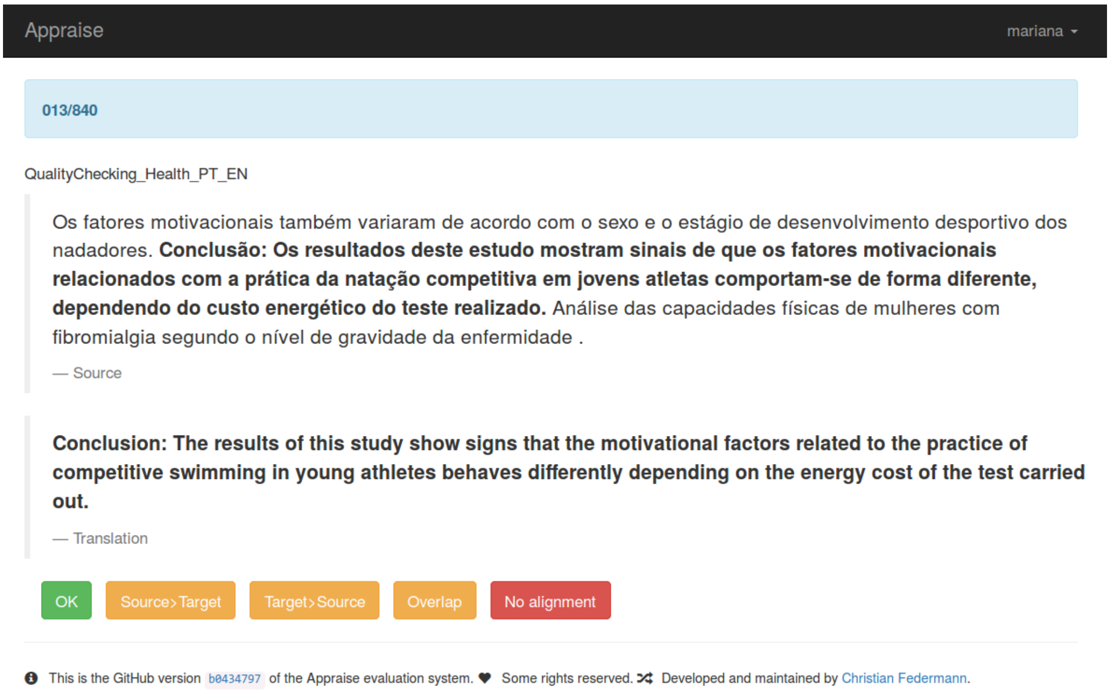
[7] http://www.scielo.mec.pt/scielo.php?script=sci_abstract&pid=S0874-48902010000300006&lng=pt&nrm=iso&tlng=pt

[8] https://opennlp.apache.org/

[9] http://nlp.cs.nyu.edu/GMA/

[10] https://github.com/cfedermann/Appraise

[11] http://www.statmt.org/moses



QualityChecking_Health_PT_EN

Os fatores motivacionais também variaram de acordo com o sexo e o estágio de desenvolvimento desportivo dos nadadores. **Conclusão: Os resultados deste estudo mostram sinais de que os fatores motivacionais relacionados com a prática da natação competitiva em jovens atletas comportam-se de forma diferente, dependendo do custo energético do teste realizado.** Análise das capacidades físicas de mulheres com fibromialgia segundo o nível de gravidade da enfermidade .

— Source

Conclusion: The results of this study show signs that the motivational factors related to the practice of competitive swimming in young athletes behaves differently depending on the energy cost of the test carried out.

— Translation

This is the GitHub version b0434797 of the Appraise evaluation system. Some rights reserved. Developed and maintained by Christian Federmann.

Figure 2: Screen-shot of the Appraise tool during the validation of the alignment of a EN/PT pair of sentences.

4. Results and Evaluation

In this section, we present the statistics on our corpus and its evaluation regarding two aspects: the quality of sentence alignment and the corpus suitability for training and evaluation of MT algorithms.

4.1. Corpus Statistics

Table 1 presents statistics of the training data, including the proportion of titles and abstracts and the total number of documents, sentences and tokens. The corpus is currently available for download[12] in the BioC XML format (Comeau et al., 2013), a format which is becoming a standard in the biomedical natural language processing (BioNLP) community. Using this format also ensures the integration of our corpus with tools and other corpora as well as making use of any of the available BioC implementations (e.g. Java, Python or C++). The following metadata are available for each document in the corpus: subject (Biological Sciences or Health Sciences), language (EN, FR, PT, ES), and zone in the document (title or abstract). The documents are split by sentences according to the analysis we obtained using the HANA database.

As discussed above, the training data is currently being used in the scope of the biomedical task in the WMT16 challenge. Besides the training data, we also released a parallel corpus of MEDLINE titles, similar to the dataset used in our previous work (Jimeno Yepes et al., 2013), the monolingual documents obtained from Scielo and all alignment output on the sentence and word level that we obtained from the GMA tool.

As observed in table 1, and due to the focus of the Scielo database on journals from Latin America, the number of documents is much larger for Portuguese and Spanish compared to French, for both categories and for both the parallel and monoligual datasets. Indeed, the number of parallel documents for FR/EN and the Biological Sciences category was so low that we do no provide any training and test datasets for it. Alternatively, it is possible to train a MT system for the Biological Sciences using documents from the Health Sciences, or even completely ignore categories and use a single system trained on the whole dataset for a given language pair.

Regarding percentages of titles and documents in the training data, table 1 shows that more abstracts are available in comparison to titles. This aspect is due to the high number of documents in Scielo that have their abstract translated to other language but not their titles, such as document S0874-48902010000300006 cited above. This is certainly a good feature of our corpus, given that previous parallel corpora of biomedical publication were restricted to MEDLINE titles (Kors et al., 2015; Jimeno Yepes et al., 2013). On the other hand, the monolingual datasets are mainly composed of titles, due to the same reason stated above, i.e., the existence of many articles whose titles were been translated to other languages. Finally, we officially released only parallel datasets that include English in the language pair. However, there are some documents which are available for other language pairs, such as ES/PT, ES/FR and FR/PT, as illustrated in Figure 3.

4.2. Alignment Quality

We manually validated the alignment at the sentence level for 200 documents (titles and abstract) for each language pair, which constitute a total of 822, 820, 607 sentences

[12] https://drive.google.com/folderview?id=0B3UxRWA52hBja0t2azlkN3d2elk&usp=drive_web

Table 1: Statistics on the parallel training and test collections according to the categories and the available languages. "T" corresponds to percentage of titles and "A" to percentage of abstracts, separated by a slash. "Docs" to total number of documents, "Lang" identifies the language,"Sents" to total number of sentences and "Tokens" to total number of tokens.

Train	Docs	T/A	Lang	Sents	Tokens
Biological Sciences					
EN/ES	17,672	49.4/97.7	EN	138,073	3,819,190
			ES	128,894	3,887,818
EN/PT	18,180	31.1/96.1	EN	128,357	3,807,296
			PT	125,717	3,598,618
Health Sciences					
EN/ES	75,856	55.6/99.5	EN	628,966	15,978,198
			ES	606,231	17,168,994
EN/PT	65,659	74.0/92.8	EN	541,272	14,457,939
			PT	525,721	14,447,017
EN/FR	1,135	64.5/99.7	EN	9,393	250,907
			FR	9,501	320,132

Test	Docs	T/A	Lang	Sents	Tokens
Biological Sciences					
EN2ES	500	100/100	EN	4,344	116,388
			ES	4,070	125,491
ES2EN	500	100/100	ES	4,113	124,343
			EN	4,405	115,045
EN2PT	500	100/100	EN	4,333	114,705
			PT	4,205	120,591
PT2EN	500	100/100	PT	4,029	114,970
			EN	4,164	108,120
Health Sciences					
EN2FR	500	100/100	EN	5,093	137,321
			FR	5,782	208,795
FR2EN	500	100/100	FR	5,784	206,559
			EN	5,178	137,638
EN2ES	500	100/100	EN	5,111	127,112
			ES	5,027	141,473
ES2EN	500	100/100	ES	5,198	144,666
			EN	5,276	128,742
EN2PT	500	100/100	EN	3,858	99,001
			PT	3,776	101,991
PT2EN	500	100/100	PT	3,826	106,735
			EN	3,930	102,813

Table 2: Statistics on the monolingual collections according to the categories and the available languages. "T" corresponds to percentage of titles and "A" to percentage of abstracts, separated by a slash. "Docs" to total number of documents, "Sents" to total number of sentences and "Tokens" to total number of tokens.

Mono	Docs	T/A	Sents	Tokens
Biological Sciences				
ES	1,007	98.4/3.5	1,248	225,904
PT	340	100/0	352	65,031
EN	24,006	95/20.6	55,346	4,897,988
Health Sciences				
ES	3,941	96.8/46.4	5,163	864,549
PT	1,466	97/11.6	2,970	298,056
FR	9	100/0	9	1,595
EN	38,214	96/10.3	68,992	988,905

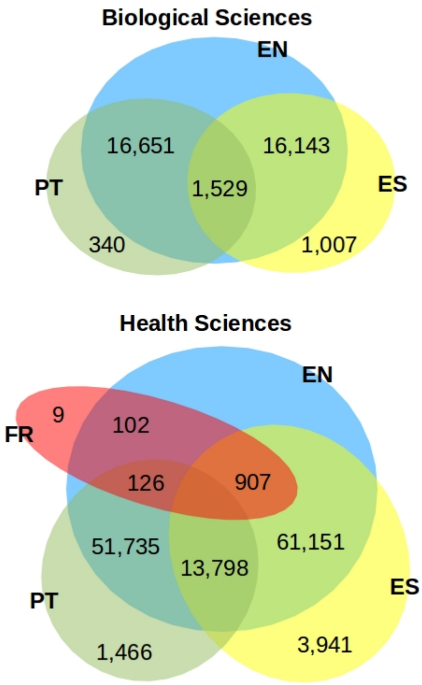


Figure 3: Size of the training datasets for each category, language pair, across language pairs and monolingual corpora. (The figure is not in scale.)

(ES/EN, PT/EN, FR/EN) for Biological Sciences and 844, 840, 977 (ES/EN, PT/EN, FR/EN) for Health Sciences. The rate of aligned sentences (i.e., "OK") varies from 79% to 85% while the rate of sentences that were not aligned (i.e., "No alignment") is less that 3%. The rate of sentences with overlap of information is also very low, up to 3.5%, while the proportion of sentences in which either the source or the target contains more information than the other language varies across the languages and subjects, but usually ranges from 5% to 10% (cf. Figure 4).

Differences between the parallel sentences occur for a variety of reasons. When the source text contains more information than the target text, or vice-versa, it is usually due to the authors failing to include some content during the translation, or adding more content instead. For instance in the English sentence *"Analysis of the composition of Brazilian propolis extracts by chromatography and evaluation of their in vitro activity against gram-positive bacteria."* contains only one word more (*"gram-positive"*) than its parallel sentence in Portuguese. Other examples exhibit more significant content difference, such as in the sentence *"Salicylic acid degradation from aqueous solutions using Pseudomonas fluorescens HK44: parameters studies and application tools."*, where all the text after the colon is missing in the Portuguese sentence.

Though the automatic alignment produced by the GMA tool has a high quality, there are still some cases in which two unrelated sentences were aligned. This is the case of the English sentence *"This richness was similar to that previously reported for the Tucumán Province, although species occurring in both provinces were mostly different."* which was aligned to the Spanish sentence *"Los mayores valores de riqueza específica se encontraron en el extremo norte de la provincia (R=32), donde a su vez se registra la mayor cantidad de muestreos."*. Nevertheless, both sentences share some words in common, such as the words *"province/provincia"* and *"richness/riqueza"*.

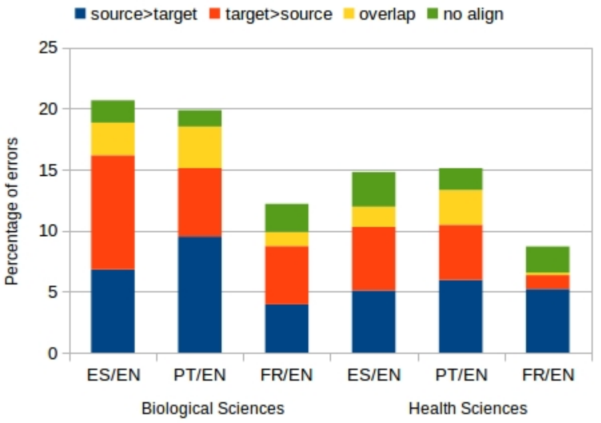


Figure 4: Percentage of the errors in the quality assessment of the corpus.

4.3. Machine Translation Experiments

The data sets developed for each language pair and science field are split to train and tune the Moses MT tool. From the whole set of aligned sentences, 10,000 sentences were kept for tuning while most of the aligned sentences were used for training. The EN/FR Health Sciences Scielo set has less than 10,000 aligned sentences, so all the sentences from this set were used for both training and tuning steps. For training Moses, we enhanced the Scielo sentences using bilingual article titles from PubMed, which helps increasing the lexical coverage of the Scielo corpus.
We followed the training example from Moses web site[13] and the results are presented in Table 3. Compared to previous work (Jimeno Yepes et al., 2013), the BLEU score of EN/ES is higher compared to a corpus using aligned bilingual abstracts. BLEU score for EN/FR is lower compared to previous work, however, the Scielo set for EN/FR is much smaller. No previous results for a similar work are available for EN/PT, but results are similar to the ones obtained for EN/ES. Our results indicate that the Scielo corpus can be effectively used to train a statistical MT system for EN/ES and EN/PT language sets.

Table 3: Baseline results for MT based on the Scielo corpus

Datasets		BLEU	
		Biological	Health
ES-EN	ES2EN	30.53	28.66
	EN2ES	32.75	31.11
PT-EN	PT2EN	29.40	31.78
	EN2PT	31.98	33.37
FR-EN	FR2EN	-	13.52
	EN2FR	-	15.34

4.4. Corpus quality

It can be argued that the evaluation of sentence alignment and the BLEU scores obtained in the machine translation experiments provide a positive evaluation of the corpus quality as a whole. However, in our manual review of sentence alignment, we found some examples where the language quality of the corpus was lacking. Minor issues include consistency in the encoding so that special characters such as apostrophes or accented letters need to be accounted for carefully. More severe issues were found with the fluency and correctness of one or both of the languages of the document perused. We believe this to be a direct result of the training of the article authors who are not professional translators, and who may have limited proficiency in one or more of the languages they need to use when writing articles. Below is a description of the error types commonly found:

- mistranslation conveying erroneous meaning: the sentence *"Certains patients ont affirmé qu'ils n'avaient aucune raison de demander des explications."* was translated by *"Some inpatients stated that they had no reason for not seeking clarity."* instead of *"...for seeking clarity."*.

- grammatical error: *"une consommation nocif"* instead of *"une consommation nocive"* (erroneous gender agreement between adjective and noun).

- lexical error: *"success in attaining vaginal delivery"* was translated by *"la réussite d'un accouchement par voie vaginale"* instead of *"la réalisation..."*.

- A combination of error types: *"All this resulted in laws favouring parents interests over embryo's rights."* was translated into French as *"Tout çela est traduit en législations qu'ont superposés l'intérêt des parents sur les droits de l'embryon."*, a sentence that conveys a different meaning from the original sentence and exhibits grammatical errors (*"traduit en"* instead of *"traduit par"*, *"qu'ont"* instead of *"qui ont"*) and erroneous lexical choices (*"superposé"* instead of *"privilégié"*). A better translation would have been: *"Tout çela s'est traduit par des législations qui ont privilégié l'intérêt des parents plutôt que les droits de l'embryon."*

Overall, these issues were found in a small number of documents in the sample used to evaluate sentence alignment. In future work, it would be interesting to assess the prevalence of these types of errors systematically.

[13] http://www.statmt.org/moses/?n=Moses.Baseline

5. Conclusions and Future Work

We presented the development of the first parallel corpus of biomedical titles and abstracts freely available in three language pairs: EN/ES, EN/FR and EN/PT. We describe the crawling of the documents from Scielo, language recognition, sentence splitting and sentence alignment using state-of-the-art tools. For quality assurance, sentence alignment were manually validated, and found to be correct for around 80% of sentence pairs in ES/EN and PT/EN in the Biological Sciences set and around 87% in FR/EN. This percentage is higher, around 85%, in the Health Sciences for ES/EN and PT/EN. Baseline machine translation experiments were performed and achieved a BLEU score higher compared to previous work for well covered language pairs.

As further work, we could evaluate the contribution of additional, in and out of domain, available corpora to improve MT results. We plan to make the training set available in community challenges (e.g. ACL WMT'16) so the research community can experiment with additional translation methods. Finally, it would be interesting to use this corpus in a task that could be used in a practical context.

6. Acknowledgments

MN acknowledges computing resources from the HPI Future SOC Lab. AN was supported by the French National Research Agency under grant CABerNeT ANR-13-JS02-0009-01.

7. Bibliographical References

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Comeau, D. C., Islamaj Doğan, R., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T. C., Wu, C. H., and Wilbur, W. J. (2013). Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Jimeno Yepes, A., Prieur-Gaston, E., and Neveol, A. (2013). Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):146.

Kirchhoff, K., Turner, A. M., Axelrod, A., and Saavedra, F. (2011). Application of statistical machine translation to public health information: a feasibility study. *Journal of the American Medical Informatics Association*, 18(4):473–478.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

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.

Névéol, A., Max, A., Ivanishcheva, Y., Ravaud, P., Zweigenbaum, P., and Yvon, F. (2013). Statistical machine translation of systematic reviews into French. In Rita Temmerman et al., editors, *Proceedings of the TIA Workshop on "Optimizing understanding in multilingual hospital encounters"*, page 4p, October 30th, 2013.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G., Kelly, L., Leveling, J., Mareček, D., Novák, M., Popel, M., Rosa, R., Tamchyna, A., and Urešová, Z. (2014). Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artif Intell Med*, 61(3):165–85, Jul.

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. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari, et al., editors, *LREC*, pages 2214–2218. European Language Resources Association (ELRA).

Wu, C., Xia, F., Deleger, L., and Solti, I. (2011). Statistical machine translation for biomedical text: Are we there yet? *AMIA Annual Symposium Proceedings*, pages 1290–1299.

8. Language Resource References

(2010). *Appraise*.

(2010). *GMA (Geometric Mapping and Alignment)*.

(2010). *Moses Phrase-based Statistical Machine Translation system*.

Neves, Mariana and Jimeno Yepes, Antonio and Névéol, Aurélie. (2016). *Scielo corpus*. ISLRN 931-667-076-873-7.