A parallel collection of clinical trials in Portuguese and English

Mariana Neves
Hasso Plattner Institute at University of Potsdam
August Bebel Strasse 88, Potsdam 14482 Germany
mariana.neves@hpi.de

Abstract

Parallel collections of documents are crucial resources for training and evaluating machine translation (MT) systems. Even though large collections are available for certain domains and language pairs, these are still scarce in the biomedical domain. We developed a parallel corpus of clinical trials in Portuguese and English. The documents are derived from the Brazilian Clinical Trials Registry and the corpus currently contains a total of 1188 documents. In this paper, we describe the corpus construction and discuss the quality of the translation and the sentence alignment that we obtained.

1 Introduction

It is well know that parallel collections of documents are valuable resources for training, tuning and evaluating machine translation (MT) tools. These are an alternative to relying on expensive bilingual dictionaries. However, parallel documents are only available for some particular languages and domains, e.g, (Koehn, 2005). Additionally, building such a corpus usually requires manual translation of documents from one language to another, which is an expensive and time-consuming task.

Even though many corpora are available for a variety of domains and languages (e.g., news text[1]), these are still scarce for biomedicine. However, domain-specific documents are indeed necessary in order to address the complexity and variety of the biomedical terminology.

Most of medical documents cannot be made freely available due to privacy issues, as it is the case of discharge summaries. Furthermore, many of such documents are only available in one language. On the other hand, scientific publications are a rich source of biomedical terminology, but these are mostly available only in the English language. Even though there has been previous work on biomedical MT using titles and abstracts of scientific publications (Jimeno Yepes et al., 2013; Wu Cuijun et al., 2011), few document collections are currently available for training MT systems. As far as we know, there are two comprehensive collections for parallel documents to support biomedical MT: (i) the UFAL Medical corpus[2] that has a focus on medicine and gathers documents derived from three research projects (KConnect, Khresmoi and HimL); and (ii) the Scielo corpus (Neves et al., 2016), which includes comparable scientific publications from a Latin American database. Both collections have supported previous MT challenges (Bojar et al., 2014, 2016).

Clinical trials are important source of information of the biomedical terminology and could be used to support training of MT systems. Such documents are the standard procedures to evaluate the effectiveness of a treatment, therapy or medication for a particular disease or ailment[3]. The aim of these documents is to recruit patients to take part on the studies, usually though invitation from the physicians. Therefore, they are usually publicly available in order to increase their visibility, for instance, in the ClinicalTrials.org database[4]. Clinical trial documents usually include information about the purpose of the trial, details of the procedure, conditions that the patient should meet, i.e., inclusion and exclusion criteria, as well as pri-

[1] http://www.statmt.org/wmt17/translation-task.html

[2] https://ufal.mff.cuni.cz/ufal_medical_corpus

[3] https://www.nhlbi.nih.gov/studies/clinicaltrials

[4] http://clinicaltrials.gov

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 36–40,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

mary or secondary outcomes. Nevertheless, most clinical trials seem to be available in only one language, which undermine their use for MT systems.

We present the first parallel corpus of clinical trials. The documents are derived from the Brazilian Clinical Trials Registry (Registro Brasileiro de Ensaios Clìnicos - ReBEC)[5]. The database currently contains 1314 registered trials (as of April 21, 2017). Documents in ReBEC are composed of many fields, such as the scientific title, the description of the intervention, inclusion criteria, exclusion criteria, primary outcomes and secondary outcomes (cf. Figure 1). For all documents, most of these fields are available in English and Portuguese and translation has probably been carried out by the responsible of the trial. The trials can be easily downloaded from the web site and are allowed be redistributed (confirmed by personal communication via e-mail).

We describe the construction of our corpus, which included parsing the XML files and performing sentence splitting, tokenization, automatic sentence alignment and manual checking the aligned sentences. We compiled a total of 1188 parallel documents and we believe that this resource can support training, testing or tuning MT systems. The documents are available at https://github.com/biomedical-translation-corpora/rebec. Given the scarce number of biomedical resources for MT, additional data is of much value in the field.

2 Corpus construction

In this section, we describe the procedure to create a parallel corpus of clinical trials. Our workflow was inspired in the one carried out for the Scielo corpus (Neves et al., 2016), even though we used different NLP components and skipped the crawling step, which is not necessary in ReBEC.

Data download. Users can easily download clinical trials from ReBEC by simply selecting some clinical trials from a list and by clicking on the check-box. It is possible to select all trials on the page by clicking on the corresponding check-box. Selected trials are then exported to a OpenTrials XML file. The only limitation is that up to ten trials are presented per page. Therefore, we had to repeat this procedure many times until we had downloaded their totality (120 files as of January 4th). We did not distinguish between the many types or topics in the trials, in order to obtain a dataset as general-purpose as possible.

OpenTrials XML Parsing. We parsed the OpenTrials XML using some procedures developed in Java. We considered only the following eight fields when parsing the XML file: (a) the trial identifier (element "trial_id"); (b) the public tittle of the trial (element "public_title"); (c) the scientific title of the trial (element "scientific_title"); (d) the interventions to be carried out in the trial (element "interventions"); (e) the inclusion criteria for taking part in the trial (element "inclusion_criteria"); (f) the exclusion criteria for not participating in the trial (element "exclusion_criteria"); (g) the primary outcome of the trial (element "primary_outcome"); (h) the secondary outcome of the trial (element "secondary_outcome").

The identification of the language is not straightforward in the OpenTrials XML format. For some fields, it is identified by the attribute "language" or "lang" in some tags, and sometimes by specific tags, such as "translation" or "outcome_translation". Nevertheless, it is always possible to identify the language of the text in each field, and therefore, it is not necessary to make use of language recognition tools.

We exported the above fields into the the BioC format (Comeau et al., 2013), a standard XML format in the biomedical NLP community. This XML format contains one "passage" tag for each of the above fields, while the name of the field and the language are informed using the so-called "infons" in the BioC format (cf. Figure 2). We tried to position the passages in the same order as they occur in the XML format in order to reduce possible errors in the automatic alignment step (cf. below) and we followed the same notation defined in the Scielo corpus (Neves et al., 2016). We obtained a total of 1188 documents.

Sentence splitting. This step consists on splitting the sentences in each of the passages, i.e., each of the fields of the trials. This is a necessary step for later align the documents sentence by sentence. We used the OpenNLP[6] tool for sentence splitting and utilized the corresponding models for English and Portuguese.

[5] http://www.ensaiosclinicos.gov.br/

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

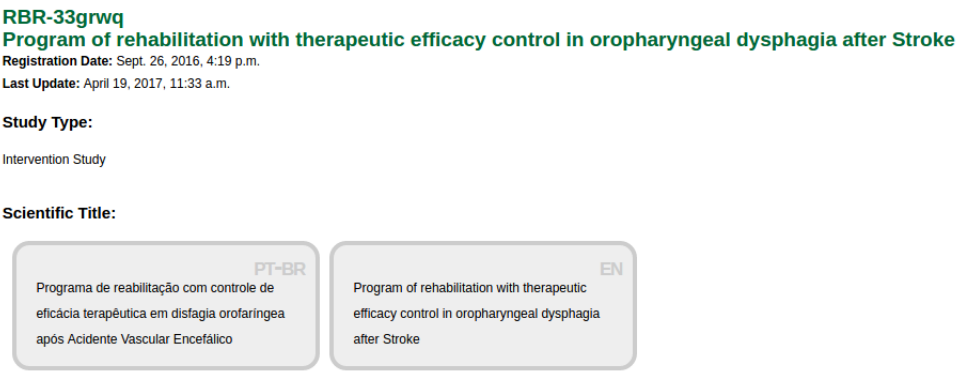

Figure 1: Screen-shot of a clinical trial in ReBEC.


```
−<collection>
  <source/>
  <date/>
  <key/>
 −<document>
    <id>RBR-2cxrpp</id>
   −<passage>
      <infon key="seq-section">1</infon>
      <infon key="section">public_title</infon>
      <infon key="lang">en</infon>
      <offset>-1</offset>
     −<text>
        Effect of Two Kinds of Therapy on Women with Patellofemoral
        Pain Syndrome
      </text>
    </passage>
   −<passage>
      <infon key="seq-section">2</infon>
      <infon key="section">scientific_title</infon>
      <infon key="lang">en</infon>
      <offset>-1</offset>
     −<text>
        Effects of Lumbo-pelvic Stabilization Training on Women with
        Patellofemoral Pain Syndrome
      </text>
    </passage>
```


Figure 2: Screen-shot of one of the document in the BioC XML format.

Sentence alignment. Similar to the work of (Neves et al., 2016), we aligned the sentences using the Geometric Mapping and Alignment (GMA) tool[7]. Sentence alignment is a necessary step for many MT tools (Sennrich and Volk, 2011). In this work, our aim was to align the sentences to further check the quality of the translation in the next step. Given the long length of the documents, a validation based on the whole document would not be feasible using the current available validation tools, e.g., Appraise (Federmann, 2010).

We converted each document to their .axis file format using scripts available in the GMA tool. In a next step, we aligned the sentences using the default parameters of the tool. We only had to inform a list of stopwords for each language and we use the following for Portuguese[8] and English[9].

Quality checking. We randomly selected a sample of 50 clinical trials to manually check the quality of the alignment, translation and sentence splitting and obtained a total of 891 items (pairs). We utilized the Appraise tool[10] (Federmann, 2010), which is freely available. Appraise includes various tasks to manually validate the quality of translations. We used the "Quality Checking" task which consists of showing the source sentence(s) (i.e., in Portuguese), and the corresponding aligned translation sentences (i.e. English). More than one sentence might be shown for any of the two languages depending on the output of the alignment tool. The validation was carried out by the author who is a native speaker of Brazilian Portuguese. Similar to (Neves et al., 2016), we adopted five options when checking the items, as described below:

- OK: correct text alignment, i.e., the English translation is a correct translation of the Portuguese source.
- Source>Target: there is more information in the source (Portuguese) text than in the translation (English) text.
- Target>Source: there is more information in the translation (English) text than in the source (Portuguese) text.
- Overlap: there is some overlap between both text but also information which are just present in each on of them.
- No alignment: wrong alignment of the sentences.

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

[8] http://www.linguateca.pt/chave/stopwords/chave.MF300.txt

[9] http://www.textfixer.com/tutorials/common-english-words.txt

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

Language	Sentences	Tokens
EN	23,843	625,881
PT	23,666	665,325

Table 1: Statistics on the size of the collection of parallel clinical trials.

At the end of the validation process, Appraise provides statistics for the chosen options and allows the user to export the results for further analysis.

3 Results and discussion

In this section we present statistics on the corpus and the results of the manual evaluation of a sample of documents. Table 1 shows statistics on the size of the collection of clinical trials for each language. The number of tokens is based on the OpenNLP tool for both languages using the corresponding available models. Even though the collection is much smaller that the ones available for Portuguese/English and Spanish/English in the Scielo corpus, it is larger than the the one available for French/English in the same corpus. Additionally, we have a higher number of documents than some of the collections available in the UFAL Medical corpus.

Table 2 shows the results of the validation of the sample of 50 clinical trials. A total of 67% of the items were correctly aligned, while overlaps and text in one language containing more information than in the other language were rather rare (around 4% in total). The "Target >Source" or "Source >Target" options were selected even when difference was minimal, such as in one case in which the English translation contained the expression "24-hour", which was not present in the Portuguese version. Some of these mistakes were also due to two sentences in one language being aligned to just one in the other language, while the corresponding second sentence was placed in the next alignment block, i.e., an error caused by the sentence alignment step.

However, in contrast to the results reported for the Scielo corpus, we obtained a much higher number (and percentage) of wrong alignments (the "No alignment" option). During validation, we noticed a high number of empty sentences, which is the result of empty lines in the original files. This mistake accounts for 27 of the wrong alignments, however, this is still only around 1/5 of the total errors for this type.

Result	No. items (%)
OK	597 (67.00%)
Source>Target	25 (2.81%)
Target>Source	15 (1.68%)
Overlap	4 (0.45%)
No alignment	250 (28.06%)
total	891 (100%)

Table 2: Results from the manual validation of the sample of 50 clinical trials using the Appraise tool (Quality Checking task).

Some wrong alignments were due to mistakes in the sentence splitting components. For instance, one Portuguese sentence ending on "[...] durante 45 minutos, num total de 16 sesses." was aligned to the English sentence "45 minutes, totaling 16 sessions.". The English sentence was mistakenly split before the token "45", and the rest of this sentence was placed on the previous alignment block. There is no clear reason on why the OpenNLP tool split the sentence at this particular point for the English sentence, but not for the corresponding Portuguese sentence.

Finally, many wrong alignments were probably due to errors from the GMA tool. In many cases, for no clear reason, sentences from one field were aligned to sentences from the adjoining field. Indeed, our input data to GMA does not distinguish the boundaries between the fields.

In general, the English translation is of good quality, although some lexical and grammar errors did occur. However, cases in which the English translation was particularly bad were rather rare, e.g., the sentence "Secondary outcomes are expected not".

4 Conclusions and future work

We presented the construction of the first parallel collection of clinical trials. Our document collection is not particularly small, in comparison with previous works, however, the quality of the alignment that we obtained was rather low. To overcome this problem, we believe that a better alignment could be obtained by carrying it out for each field separately, instead of the complete document. However, given that some fields appear more than once and in no particular order in the file, precisely extracting the fields is not a straightforward task. Further, we plan to try other sentence alignment tools, besides the GMA tool, and analyze the suitability of the corpus for training biomedical MT systems. Finally, our future versions of the cor-

pus will also include additional fields to the ones considered here.

Acknowledgments

We would like to thank ReBEC for granting us permission to redistribute the clinical trials.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 12–58.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation (WMT16) at the Conference of the Association of Computational Linguistics*, pages 131–198.

Donald C. Comeau, Rezarta Islamaj Doan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database* 2013.

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

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

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

Mariana Neves, Antonio Jimeno Yepes, and Aurlie Nvol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

R Sennrich and M Volk. 2011. Iterative, mt-based sentence alignment of parallel texts. In *NODALIDA 2011, Nordic Conference of Computational Linguistics*.

Wu Cuijun, Xia Fei, Deleger Louise, and Solti Imre. 2011. Statistical Machine Translation for Biomedical Text: Are We There Yet? *AMIA Annual Symposium Proceedings* 2011:1290–1299.