English-Portuguese Biomedical Translation Task Using a Genuine Phrase-Based Statistical Machine Translation Approach

José Aires[1,2] **Gabriel Pereira Lopes**[1,2] **Luís Gomes**[1,2]
[1] NOVA LINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
[2] ISTRION BOX, Translation and Revision, Lda, Portugal
{jose.aires,gabriel.lopes,luis.gomes}@istrionbox.com

Abstract

Our approach to produce translations for the ACL-2016 Biomedical Translation Task on the English-Portuguese language pair, in both directions, is described. Own preliminary tests results and final results, measured by the shared task organizers, are also presented.

1 Introduction

This paper shows how we obtained our results using our patented Machine Translation system (Lopes et al., 2015) to produce translations for the English-Portuguese language pair from the Biomedical Translation Task.

Our approach differs from common Statistical Machine Translation approaches like Moses (Koehn et al., 2007) in several aspects:

- phrases are not analyzed at their word level in any model;
- the language model depends on the target alternatives of given adjacent sources and does not try to avoid null scores to phrases that do not occur;
- the translation score is not log-linear, but instead a tuned weighted average between the translation model and the language model, and so no smoothing techniques are required;
- several models can be used with different relevances or weights; and
- instead of simply relying on statistics, we include human validation and correction on several stages of the system, namely for validating extracted term translations, to improve the quality of the source data used in the automatically produced translations.

As requested, the translation results were produced using the sentence-aligned training data described below (for the English-Portuguese language pair, in our case), provided by the shared task organizers:

- **medline-pubmed**: parallel corpora from medline;
- **scielo-gma-biological**: parallel biological documents from the Scielo database (Neves et al., 2016); and
- **scielo-gma-health**: parallel health documents from the Scielo database (Neves et al., 2016).

Table 1 shows the features of the English (en) and Portuguese (pt) languages of each provided corpora, namely their number of lines and words.

corpus	lines	words
medline-pubmed-en	74,645	917,307
medline-pubmed-pt	74,645	1,041,079
scielo-gma-biological-en	120,301	3,338,244
scielo-gma-biological-pt	120,301	3,736,817
scielo-gma-health-en	507,987	13,443,076
scielo-gma-health-pt	507,987	14,901,240

Table 1: Training corpora data after normalization.

The translation task then consisted in translating one document from English to Portuguese and another from Portuguese to English, for both the biological and the health domains, with the number of lines and words from those test documents shown in Table 2.

Besides the provided training data, we have also included our English and European Portuguese bilingual lexicon (described in §2.3.2), as well as our named entities database, for additional term coverage.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 456–462,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

document	lines	words
biological_pt2en	4,029	119,410
biological_en2pt	4,333	111,038
health_pt2en	3,826	111,073
health_en2pt	3,858	96,240

Table 2: Test documents data after normalization.

The training corpora had to undergo several processing stages in order to support the production of the intended translations, as described in the following section.

2 Data Processing

In order to produce translations, our system (like any other Statistical Machine Translation system) requires a translation model and a language model to support the translation decoding stage. To calculate such models the available data had to go through several processing steps described in the following subsections.

Since each of the training corpus has been made available separately, we also opted to process each of them separately so that we were then able to use them with different weights, assigning more or less weight to models with higher or lower relevance, respectively. See extended explanation in §4.

2.1 Considerations about the provided data

It should be noted that we have detected a few flaws in the provided data, namely several sentences incorrectly considered as parallel, as well as the existence of many spelling errors, not only in the training data, but also in the testing documents.

We believe that many of the typos result from PDF extraction and/or OCR processes, which are never perfect, having found and corrected a total of 127,198 misspellings. Yet, it should be noted that some misspelling errors are easy to correct, but errors which still produce correct words require sentence analysis which was not carried out.

Some of the parallel problems are illustrated, for instance, by having the first Portuguese line from medline-pubmed "*ERRATA.*" aligned with the first English line "*Inequalities in self-rated health: an analysis of the Brazilian and Portuguese populations.*", which should be "*ERRATA.*" instead.

Filtering wrong translation units as the one above, as well as translation units which the language was not Portuguese, reduced this corpora by almost 2,000 translation units.

Some errors were simply detected by chance, like first and last entries of medline-pubmed, while other errors were detected by looking at the untranslated terms in the initial testing §3 and realizing that some terms were misspellings, as well as spelling and vocabulary differences between European and Brazilian Portuguese.

corpus	lines	words
medline-pubmed-en	74,645	917,307
medline-pubmed-rev-en	72,651	898,051
medline-pubmed-pt	74,645	1,041,079
medline-pubmed-rev-pt	72,651	1,006,069

Table 3: medline-pubmed revision impact.

Table 3 shows the differences between the original version medline-pubmed and its revised version medline-pubmed-rev. The reduction in size towards the revised version is mainly due to the removal of non-parallel sentences.

However, efforts to correct such situations were only made over the mentioned medline-pubmed parallel document set, since the other sets were significantly larger, as shown in Table 1. Also, no corrections were applied to the testing documents because we assumed they were not supposed to be edited.

Yet, another "noise" element was the already mentioned difference in spelling and vocabulary between European Portuguese (which has been our main focus of attention throughout our research experience) and Brazilian Portuguese (the version of the provided biomedical data), which can also impact results negatively.

2.2 Text tokenization and normalization

Text tokenization ensures that words are properly separated by a single blank space, while normalization ensures that they are represented by a "standard" version. In English, this means that cases like "*wasn't*" or "*isn't*" are going to be replaced by "*was not*" and "*is not*", respectively. In Portuguese, this means that cases like "*do*" (*of the*) or "*nas*" (*in the*) are going to be replaced by "*de o*" (*of the*) and "*em as*" (*in the*), respectively. These tokenization and normalization changes are reverted when presenting the final translation results.

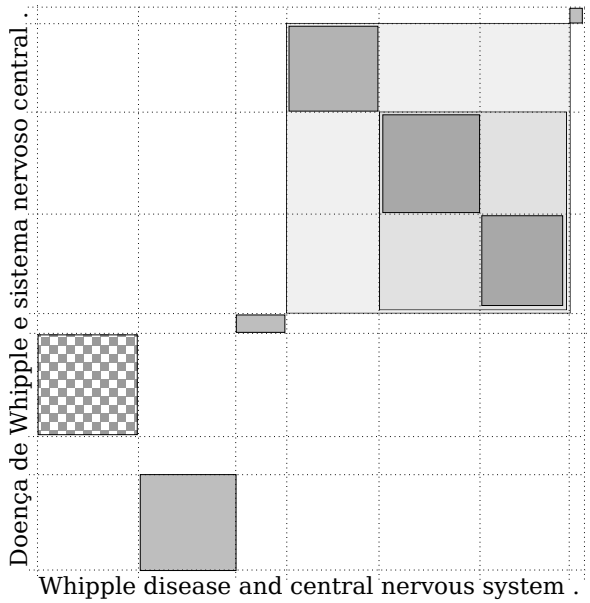

Figure 1: Example lexicon- and cognate-based alignment of a short sentence from the medline-pubmed corpus. Gray-filled rectangles represent word- and phrasal-matches from the lexicon while the checkerboard-filled rectangle shows a cognaticity-based match.

2.3 Phrase alignment

Phrase-level alignment was obtained with a modified version of the lexicon-based aligner proposed by Gomes (2009). The aligner matches bilingual phrase pairs provided in an input lexicon (described ahead in §2.3.2) and selects a maximal-coverage[1] subset of *coherent* alignments. While the original method imposed a monotonicity constraint, i.e. it selected a maximal-coverage chain of phrase alignments without allowing phrase reorderings, the new method applied has a more relaxed coherency criteria: it only requires that a source-language phrase is not simultaneously aligned with two distinct target-language phrases. Therefore, it allows phrase reordering as shown in the example in Figure 1.

2.3.1 Alignment as an optimization problem

Similar to the ILP (Integer Linear Programming) solution proposed by (DeNero and Klein, 2008), we treat the alignment problem as an optimization problem, but we employ a greedy optimization algorithm which allows us to align longer sentences with reasonable time and memory. The algorithm constructs a solution (a set of coherent alignments) incrementally. It starts by settling alignments of longer phrases, which tend to be more reliable, and progresses towards shorter phrases or words, which are allowed to align only if they are coherent with previously settled alignments.

2.3.2 Input bilingual lexicon

Our EN-PT input lexicon has 931,568 manually validated translations (words and phrases). This lexicon has been compiled in a long term effort started in the context of project ISTRION[2]. The translations were extracted automatically from several corpora, including Europarl (Koehn and Monz, 2005), JRC-Acquis (Steinberger et al., 2006), OPUS EMEA (Tiedemann, 2009) and others, using a combination of complementary alignment and extraction methods: GIZA (Och and Ney, 2003), Anymalign (Lardilleux and Lepage, 2009), spelling similarity measure SpSim (Gomes and Lopes, 2011) combined with co-occurrence Dice measure, and others. The automatically extracted word and phrasal translations were automatically classified, prior to human validation, using an SVM classifier trained on previously validated translations as described by Mahesh et al. (2015). The automatic classification speeds up human validation because very few translations (less than 5%) are incorrectly classified, and only those need to be manually labeled as correct or incorrect.

We did not perform any extraction or validation of new translations from the corpus provided for this shared task. We did, however, complement our lexicon with cognate and homograph alignments using the SpSim (Gomes and Lopes, 2011) spelling similarity measure.

2.3.3 Lexicon coverage

Our lexicon covers 59.5% of the EN corpus tokens and 55.4% of the PT corpus tokens. There were 143,317 unique phrasal translations matched out of 931,568 in our lexicon. The cognaticity-based matching was responsible for aligning 8% of the EN corpus and 7.2% of the PT corpus[3]. The remainder 32.5% of the EN corpus and 37.4% of the PT corpus were left unaligned. These unaligned tokens are handled as gaps by the phrase table extraction algorithm described in (Aires et al., 2009).

[1] *Maximal-coverage* means that the selected phrase alignments cover as much text as possible from both sentences

[2] Project ISTRION was funded by the Portuguese Foundation for Science and Technology under contract PTDC/EIA-EIA/114521/2009

[3] cognaticity alignment was applied only to tokens not covered by the input lexicon

2.4 Language model training

The language model used is supported by the indexation of the texts in each language of the provided corpora. Such indexation will support determining the likelihood of the occurrence of phrases in the target language for the several adjacent translation fragments in decoding, a process based on the structures presented in (Aires et al., 2008).

2.5 Translation model training

The translation model depends on the alignment to determine phrase translation equivalents by establishing phrase relations between source and target languages, as well as to determine a degree of likelihood of those same relations, to be used in decoding to produce new translations, a process based on the methodology presented in (Aires et al., 2009).

2.6 Decoding

The decoding stage is the one that will finally produce the actual translations. First, an original text is fragmented into smaller pieces of text, which will then be used to retrieve their corresponding translations. The several combinations of the translations of those smaller pieces will represent many possible translations and the purpose of decoding is to find the most likely one, according to the provided scores from the language and the translation models. As mentioned before, separate models can be obtained from separate corpora and be assigned with different relevances or weights, according to their importance to the translation in question.

As such, and as explained in Lopes et al. (2015), decoding is carried out as a best path finding in a directed acyclic graph, where its edges are weighed by: the translation model score between source and target phrases; and the language model scores between adjacent target phrases. Each complete path will represent a possible translation in which the final score is a composition of the scores of the several edges that compose the given path. An additional penalty is introduced to provide lower scores to larger paths, which are known to produce worse results.

3 Initial Testing Preparation

Since no development data was supplied, we took the initiative to prepare some development sets in order to have an idea of the most promising set of parameters to be used in our system over the provided data to produce the intended translations. As such, several documents were removed from the original training data, composed by the medline-pubmed, biological and health sets, applying the training methods on the remaining documents and using the selected ones to translate and compare the translations against their originals by determining their BLEU (Papineni et al., 2002) scores. However, in order to get a clearer picture of the type of results that could be expected, some additional tests were carried out including the selected set of documents in the training data.

Our translation model supports: a conservative extraction approach, which is more restrictive, allowing fewer translation equivalents, having a lower recall but a higher precision; and a flexible extraction approach, which is more permissive, allowing a larger number of equivalents but at the cost of an increase of incorrect ones. We were interested in evaluating the impact of both approaches on results.

Table 4 shows the average results on both translation directions of those preliminary tests, consisting of the average BLEU scores for the conservative (cons.) and flexible (flex.) approaches, as well as the average times taken to translate the documents on either extraction approaches. Those results concern the following configurations:

- **full**: the documents used for testing were not removed from the training set (medline-pubmed, biological and health);
- **dev**: the documents used for testing were removed form the training set;
- **dev-europarl**: the same as dev, but including the europarl corpus; and
- **dev-europarl-low**: the same as dev-europarl, but assigned a lower relevance to the europarl corpus.

configuration	cons.	flex.	time
full	83.98	81.97	15.1 s
dev	51.72	55.46	3.5 s
dev-europarl	52.34	55.98	49.9 s
dev-europarl-low	52.54	56.21	46.8 s

Table 4: Initial testing results.

These preliminary tests have shown that the flexible extraction approach produced on average better translation results when the reference documents were not included in the test set, which is the normal testing situation, so we used the flexible approach.

The Europarl corpus[4], which is significantly larger (54,543,044 words in English and 60,375,477 words in Portuguese), was tested as a source of additional term coverage, which allowed a translation quality improvement lower than 1 BLEU point. However, given its significant increase in processing time because of its large size, a time increase around 14 times larger, we had to drop it from the submission tests due to deadline constraints. Additionally, these results show that assigning a lower relevance to a corpus from a totally different domain may have some positive impact on average results.

Once we have decided, from this initial testing preparation, which would be the most promising and interesting features to use in the final runs, we ran the training processes again to include the documents that have been left out, this way using the full data provided by the organizers for the runs to be submitted.

4 Submitted Results

Considering that the test documents to be translated, provided by the shared task organization, share their domain with the training data, we decided to propose for submission the three possible translation runs for each document according to the criteria described in each of the following subsections.

4.1 Run 1

This run uses the medline-pubmed, biological and health training corpora with the same relevance to translate every translation test document. These can be considered our simplest set of tests since the possible model relevance difference is not explored and no additional sources are included. In this case we achieved a total of 7228 unique untranslated terms[5].

4.2 Run 2

This run also uses the medline-pubmed, biological and health training corpora, but assigns a higher relevance to the biological corpora to translate the biological test documents and then assigns a higher relevance to the health corpora to translate the health test documents. Because the changes introduced in this set of tests only concerned the relevance of the models, the total of 7228 unique untranslated terms did not change.

4.3 Run 3

This last run shares the same features as the previous run (assigning higher relevances to corresponding corpora) but this time our bilingual lexicon and named entities database was included for term coverage improvement, and an alignment based on cognates (Gomes and Lopes, 2011) is used.

About our bilingual lexicon, considering that it was built mainly from the European legislation, it was given a lower relevance because past experiences have shown us that, when the domain is not shared with the texts to be translated, it should not have the same relevance in order to reduce the probability of using inadequate terms for the intended translation domain or subject. Again, this is a situation that has also been confirmed and noted in Table 4 between dev-europarl and dev-europarl-low: reducing the relevance of europarl contributed to a slight score increase compared to when the relevance is the same.

As a side note, translating the tests took nearly 14 hours for each run[6]. Had we included europarl, judging by Table 4, we would have taken nearly 200 hours, which is more than a week, expecting to simply gain 0.75 BLEU points, on average, so we had no other option than leaving it out. Such increase in translation time is due to the substantial increase of translation equivalents available for decoding from such a large corpus.

The decision to carry out the alignment based on cognates was taken because after a first run of tests we realized that many of the untranslated terms referred to medical terms and diseases, which shared many letters between both languages and therefore had a high level of cognaticity.

All these changes allowed a significant reduction of the unique untranslated terms to a total of 4700, and for all the reasons in this subsection, we have considered this run as being our best.

[4] http://www.statmt.org/europarl/

[5] Terms can have one or more words

[6] On a 3.3GHz CPU with 32GB RAM and 4TB disk

5 Conclusions and Future Work

The scores of our submitted translations are shown in Table 5.

run	score
Istrionbox_run1_biological_en2pt	17.55
Istrionbox_run2_biological_en2pt	16.47
Istrionbox_run3_biological_en2pt	16.45
Average	**16.80**
Istrionbox_run1_biological_pt2en	20.88
Istrionbox_run2_biological_pt2en	20.17
Istrionbox_run3_biological_pt2en	20.14
Average	**20.40**
Istrionbox_run1_health_en2pt	19.01
Istrionbox_run2_health_en2pt	18.33
Istrionbox_run3_health_en2pt	18.37
Average	**18.57**
Istrionbox_run1_health_pt2en	21.50
Istrionbox_run2_health_pt2en	20.17
Istrionbox_run3_health_pt2en	20.62
Average	**20.76**

Table 5: Initial testing results.

The results obtained were clearly below what we had expected. And what is most disturbing is the negative impact of features we expected to improve results, an expectation backed by our own tests.

However, there are a few reasons we can think of for these values, namely the way the BLEU measure has been calculated (case sensitivity and synonyms penalty - translating "home" instead of "house" might be perfectly fine), the differences between European Portuguese and Brazilian Portuguese, and the presence of several spelling and alignment errors in the training data.

Nonetheless, we can still take several actions to improve our system: namely testing both parallel corpora, health and biology, with identical weights: using Europarl and eventually EMEA corpus; the refinement of our phrase translation extraction; the extraction of specific bilingual terminology, additionally to the use of cognaticity; subsentence realignment after the bilingual terminology extraction, and a more efficient implementation of the patterns (comparable to a hierarchical translation) application.

Acknowledgments

This work was supported by ISTRION BOX, Fundação para a Ciência e Tecnologia through research project ISTRION (contract PTDC/EIA-EIA/114521/2009), individual PhD grants SFRH/BD/48839/2008, SFRH/BD/65059/2009, SFRH/BD/64371/2009, and NOVA LINCS (ref. UID/CEC/04516/2013). We would also like to thank Hugo Delgado for his support.

References

J. Aires, G. P. Lopes, and J. F. da Silva. 2008. Efficient multi-word expressions extractor using suffix arrays and related structures. pages 1–8. CIKM-ACM.

J. Aires, G. P. Lopes, and L. Gomes. 2009. Phrase translation extraction from aligned parallel corpora using suffix arrays and related structures. In *Progress in Artificial Intelligence*, volume 5816 of *LNAI*, pages 587–597. Springer-Verlag Berlin Heidelberg.

J. DeNero and D. Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of ACL-08: HLT, Short Papers*, pages 25–28, Columbus, Ohio, June. ACL.

L. Gomes and G. P. Lopes. 2011. Measuring spelling similarity for cognate identification. In *Progress in Artificial Intelligence*, volume 7026 of *LNAI*, pages 624–633, Lisbon, Portugal, October. Springer.

L. Gomes. 2009. Parallel texts alignment. Master's thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Monte de Caparica, Portugal.

P. Koehn and C. Monz. 2005. Shared task: Statistical machine translation between european languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124. ACL.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. ACL.

A.n Lardilleux and Y. Lepage. 2009. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing*, pages 214–218, Borovets Bulgaria, 09.

G. P. Lopes, J. Aires, and L. Gomes. 2015. Statistical machine translation computer system and method. Submitted at National (Portugal) Level (INPI), 8. Provisional Patent Request No. 0151000065353.

K. Mahesh, L. Gomes, J. Aires, and G. P. Lopes. 2015. Selecting translation candidates for parallel corpora alignment. In *Progress in Artificial Intelligence*, volume 9273 of *LNAI*, pages 723–734, Coimbra, Portugal, September. Springer.

M. Neves, A. J. Yepes, and A. Névéol. 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)*, Paris, France, may. European Language Resources Association (ELRA).

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

K. Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, ACL '02, pages 311–318, Stroudsburg, PA, USA. ACL.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC'2006 pp. 2142-2147. Genoa, Italy, 24-26 May 2006*, Genoa, Italy, 5. ELRA.

J. Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.