Translation of Biomedical Documents with Focus on Spanish-English

Mirela-Stefania Duma and Wolfgang Menzel
University of Hamburg
Natural Language Systems Division
{mduma, menzel}@informatik.uni-hamburg.de

Abstract

For the WMT 2018 shared task of translating documents pertaining to the Biomedical domain, we developed a scoring formula that uses an unsophisticated and effective method of weighting term frequencies and was integrated in a data selection pipeline. The method was applied on five language pairs and it performed best on Portuguese-English, where a BLEU score of 41.84 placed it third out of seven runs submitted by three institutions. In this paper, we describe our method and results with a special focus on Spanish-English where we compare it against a state-of-the-art method. Our contribution to the task lies in introducing a fast, unsupervised method for selecting domain-specific data for training models which obtain good results using only 10% of the general domain data.

1 Introduction

The 2018 Biomedical Translation Task, held as part of the Third Conference on Machine Translation, aims at evaluating systems on scientific publications from Medline (Neves et al., 2018). The task is particularly challenging as there is still not enough bilingual medical data available for training high quality Machine Translation (MT) systems. We develop and apply a data selection method on five out of the nine language pairs addressed by the task: English-Spanish, Spanish-English, English-Portuguese, Portuguese-English and English-Romanian.

Data selection, as a domain adaptation technique, exploits all available (bilingual) general domain corpora with the purpose of extracting sentences that have a strong relationship to a given in-domain. All sentences from the general domain pool are scored according to a similarity function/ algorithm/ method and after being sorted, the most similar ones are selected to take part in the MT training pipeline. The subsampling is usually done using a threshold, which is the number of sentence (pairs) or a percentage of the sentences to be considered in-domain.

We introduce a data selection method which is fast to apply and yields good results when compared with a strong baseline and a state-of-the-art method. The simplicity of the method has at its core term frequencies and a newly developed similarity function. On the one hand, no models need to be trained and the method is unsupervised, but on the other hand, the method does not consider the context of the words or their semantics. However, the results are very encouraging with BLEU (Papineni et al., 2002) scores between 31.05 and 41.84 for four language pairs.

The paper is structured as follows: the next section briefly presents related work, Section 3 describes the experimental results along with a description of our algorithm, Section 4 gives an overview of the results obtained in the task and additional experiments and the last section presents conclusions and future work.

2 Related work

Related work in data selection is ample, therefore this section only mentions methods that fit in the same category with our method and we also shortly describe the widely known state-of-the art method of performing data selection, introduced by Axelrod et al. (2011), since it is the chosen method for comparing results in this paper.

Our scoring function relies heavily on term frequency. Therefore, it falls in the category of TF-IDF[1] based approaches. Hildebrand et al. (2005) uses TF-IDF to produce vector representations of sentences. Then the cosine of the angle between the sentence vectors is interpreted as the similar-

[1]Term Frequency - Inverse Document Frequency

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 637–643

https://doi.org/10.18653/v1/W18-64071

ity between the sentences. A similar approach is given in Eck et al. (2005) where a weighting scheme based on TF-IDF by means of unseen n-grams and sentence length is applied and cosine is also used as means of determining sentence similarities. In contrast to these methods, we use only the term frequency in computing our similarity scores and we make no use of the cosine. Instead, we focus on the relative difference between a term that appears in the general domain and in the in-domain and simply multiply it by a weighting scheme that has empirically proved to be effective. Our method is also related to the other methods from the TF-IDF category with respect to its simplicity.

To compare our results with other approaches we apply the modified Moore-Lewis method which is based on (Moore and Lewis, 2010): given the source side of an in-domain corpus and a random subsample of the source side of a general domain corpus, a language model (LM) is trained on each one of them. The sentences from the general domain are scored by the difference of the cross-entropy of a sentence according to the in-domain LM and the cross-entropy of the same sentence according to the general domain LM. Axelrod et al. (2011) modified the scoring by applying the same procedure also to the target side of the corpora and afterwards summing the scores. We refer to this method as *MML (modified Moore-Lewis)* in the rest of the paper.

3 Experiments

This section describes the experimental settings including the corpora and the tools used, as well as the data selection algorithm we developed.

3.1 Corpora

The general domain data consisted of a concatenation of the Commoncrawl[2] corpora and the Wikipedia (Wolk and Marasek, 2014) corpora for English-Spanish and Spanish-English, Paracrawl[3] and Wikipedia for English-Portuguese and Portuguese-English and Paracrawl for English-Romanian. For the in-domain, we used the EMEA (Tiedemann, 2012) corpora for all language pairs and the Scielo corpora (health and biological) provided by the WMT 2016 Biomedical task (Neves et al., 2016) for all language pairs except for English-Romanian where Scielo training data was not available.

The development set for the English-Spanish and Spanish-English experiments was a concatenation of the Khreshmoi development set from the Medical Task of WMT 2014[4] and the ECDC corpus made available by UFAL[5]. The motivation for using a concatenation of two medical development sets is that we aimed at diversity in the medical data. Even though ECDC is a very small corpus consisting of only 2357 sentence pairs (for English-Spanish), combining it with Khreshmoi (500 sentence pairs) would have resulted in a quite big development set which would have made the tuning of the SMT systems very time and memory intensive. Therefore, we applied a cleaning step to ECDC which meant limiting the size of the sentences to a minimum of 20 words and a maximum of 80 words. After applying this preprocessing step, the ECDC set was down to 850 sentences, resulting in a total development set of 1350 sentences. For the experiments involving Portuguese, a sample of 1000 sentences from the Scielo development set from WMT 2016[6] was used for tuning purposes. As for the Romanian experiments, also a sample of 1000 sentences was used, but from the ECDC corpus.

Statistics including the number of sentences after preprocessing for every corpus used for the training of the MT systems is given in Table 1.

Track / Corpora	EN-ES	EN-PT	EN-RO
Commoncrawl	1.8M	-	-
Paracrawl	-	2.1M	2.4M
Wikipedia	1.6M	1.6M	-
EMEA	678K	1.08M	994K
Scielo-gma 2016	166K	613K	-

Table 1: Corpora used for DSTF

3.2 Tools

For text processing we used the *nltk* toolkit(Bird et al., 2009), the WordNet (Fellbaum, 1998) lemmatizer for English and the Snowball stemmer (F. Porter, 2001) for Spanish, Portuguese and Romanian.

The SMT systems were trained using the Moses toolkit (Koehn et al., 2007) and the Experiment Management System (Koehn, 2010). The preprocessing of the data consisted in tokenization,

[2]http://commoncrawl.org/

[3]https://paracrawl.eu/index.html

[4]http://www.statmt.org/wmt14/medical-task/

[5]http://ufal.mff.cuni.cz/ufal_medical_corpus

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

Algorithm 1 DSTF Filtering

procedure PREPROCESS_CORPUS($\mathcal{C}$)
  $tokenize(\mathcal{C})$
  $lowercase(\mathcal{C})$
  $removeStopWords(\mathcal{C})$
  $lemmatize(\mathcal{C})$ ▷ or stem if unavailable
  $keepWords(\mathcal{C})$
  $wordCount(\mathcal{C})$

procedure FILTER($\mathcal{GEN}_{side}, \mathcal{IN}_{side}$) ▷ *side* refers to either source or target
  Preprocess_Corpus($\mathcal{GEN}_{side}$)
  Preprocess_Corpus($\mathcal{IN}_{side}$)
  for each sentence $s \in \mathcal{GEN}_{side}$ **do**
    for each word $w \in s$ **do**
      if $count(w, \mathcal{GEN}_{side}) = 0$ **then**
        $weight = 0$
      else
        $weight = {count(w, \mathcal{IN}_{side})}/{count(w, \mathcal{GEN}_{side})}$
      $score_w = \left(\frac{2 \cdot (count(w, \mathcal{IN}_{side}) - count(w, \mathcal{GEN}_{side}))}{count(w, \mathcal{IN}_{side}) + count(w, \mathcal{GEN}_{side})}\right)^2 \cdot weight$
    $score_s \mathrel{+}= score_w$ ▷ all intermediate scores contribute to the final score

cleaning, lowercasing and normalizing punctuation. Our language model (LM) was obtained by interpolating (Schwenk and Koehn, 2008) the LM estimated using the general domain data and the LM estimated on the in-domain data. We used the SRILM toolkit (Stolcke, 2002) and Kneser-Ney discounting (Kneser and Ney, 1995) for estimating 5-grams LMs. All the experiments benefited from the interpolated language model, including the strong baseline and the *MML* experiment. As for the chosen state-of-the-art method, *MML*, we used the implementation available from Moses.

Tuning of the systems was done with MERT (Och, 2003) and GIZA++ (Och and Ney, 2003) using the default *grow-diag-final-and* alignment symmetrization method for word alignment.

3.3 Data selection using Term Frequency

Using bag of words to represent sentences and term frequency to compute similarity became unpopular due to its limitations, namely no integration of semantic information and ignoring the context of words (Le and Mikolov, 2014). However, through the work presented here we aim at applying this straightforward method to data selection for SMT with a new weighting scheme. Our scoring algorithm builds a profile consisting of word frequencies for each domain, for the source language and the target language. To build the profile for a corpus, all of its sentences undergo a preprocessing step: tokenization, lowercasing, removal of stop words and lemmatization or stemming in the case a lemmatizer was not available for a language (procedure $Preprocess\_Corpus$). In the end, numbers or punctuation marks are ignored and only words contribute to the scoring. For word count occurrence we used the script $ngram - count$ from SRILM.

Algorithm 1 can be applied either on the source or on the target sides of the corpora. For example, when considering the source side, for every sentence from the lemmatized (or stemmed) general domain data, we iterate through all its words. Given sentence s and the word w, we square the relative difference between the term frequency of w in the in-domain profile, $count(w, \mathcal{IN}_{side})$, and the term frequency of w in the general domain profile, $count(w, \mathcal{GEN}_{side})$. We use the same relative difference formula as in (Kešelj et al., 2003) which uses character n-grams and profiles built using the most frequent character ngrams for authorship attribution. In contrast to this, we used all the words appearing in the corpora and modified the formula by introducing a weighting scheme. Note that due to the squaring, the direction of the subtraction does not matter. The difference is multiplied by a weight and the arithmetic mean of $count(w, \mathcal{IN}_{side})$ and $count(w, \mathcal{GEN}_{side})$. The weight represents the impact that w made in the

sentence and we empirically determined it. When using only the formula from Kešelj et al. (2003) adapted to our data selection task, the results are of poor quality. Our contribution to the formula lies in introducing the weighting scheme which gives much better results than the original formula. To profit from both the source and the target corpora, summing up the scores for the source language and the scores for the target language seems to be an attractive solution. We refer to our method as *DSTF (Data Selection via Term Frequency)*.

The method has a very important advantage if compared to state-of-the-art methods: scoring is very fast for a general domain corpus (on average, the scoring step took half an hour). The results are satisfactory and will be presented in the following section.

4 Results

We report the automatic evaluation results obtained in the WMT task for five language pairs and then we present further experiments for the Spanish-English language pair. BLEU was used as an evaluation metric by the WMT Biomedical organizers and in addition to BLEU we also used METEOR (Lavie and Agarwal, 2005) for further evaluating the Spanish-English experiments.

4.1 WMT Biomedical Results

Each team was allowed to submit a maximum of three runs. For every language pair that we used to evaluate our method on, we submitted three runs as follows: the first run only considers the scores obtained using the English side of the training corpora, the second run made use of only the non-English side of the training corpora and for the third run the scores for both the source and the target sides were summed up to form a single score.

The aim of data selection is to identify in the general domain pool the top $\mathcal{N}$ most similar sentences to an in-domain, where $\mathcal{N}$ is determined empirically and is usually a small number or percentage. We experimented for this paper with $\mathcal{N} = 10\%$ since the maximum of runs allowed was three and we had three variations of the method, but we intend to conduct a range of experiments with more percentage values in future work. Table 2 presents the number of sentence pairs that were subsampled along with the total number of sentence pairs that were used in the training of MT systems.

Language pair	EN-ES	EN-PT	EN-RO
10% of Gen	350K	378K	245K
total training data	1.62M	2.07M	1.24M

Table 2: Corpora used for DSTF

The BLEU results obtained using *DSTF* are encouraging: a BLEU score of 41.84 for Portuguese-English ranked our method on the third place out of seven runs submitted by three institutions. For English-Portuguese, our BLEU scores are close to 34 for all runs. The Spanish-English automatic evaluation achieved scores around 35-36 and for English-Spanish around 31. The smallest BLEU scores were measured for English-Romanian where we obtained scores close to 14. This is not surprising considering the fact that compared to the other language pairs there was less biomedical training data available. In particular, no Scielo training corpus was available although translating from English to a morphologically rich language like Romanian is considered difficult. The BLEU scores for each run are given in Table 3. We note that the differences between each run, for every language pair, are insignificant except for one language pair, therefore we conclude that either one of the algorithm variations can be successfully applied as a fast data selection technique that yields good translations (BLEU scores between 31 and 42 for four out of five language pairs).

Language pair	EN-ES	ES-EN	EN-PT	PT-EN	EN-RO
run 1	31.32	36.16	34.92	41.84	14.60
run 2	31.05	35.17	34.19	41.80	14.39
run 3	31.33	36.05	34.49	41.79	14.07

Table 3: BLEU scores reported by WMT

4.2 Spanish-English Additional Experiments

For Spanish-English, the best performing variant of our method was run 1 - using only the English side of the corpora in the algorithm. We evaluated our *DSTF-EN* method against a strong baseline (that uses an interpolated LM), a baseline trained using only the in-domain data and the state-of-the-art method *MML* for the Spanish-English language pair[7]. Following recommendations from H. Clark et al. (2011) and standard practices, we tuned the systems three times and report in Table 4 the averaged BLEU scores.

[7]Due to time limitations, we will evaluate further language pairs against *MML* in the future work.

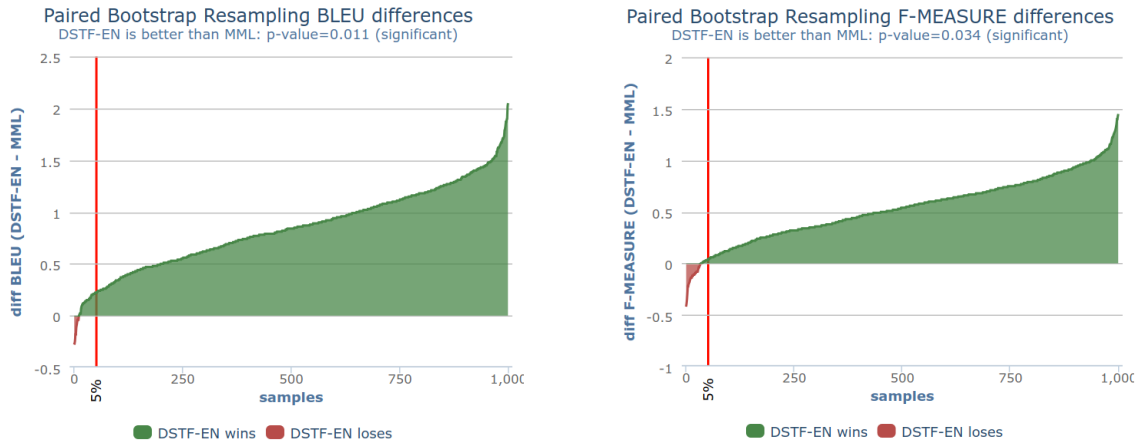

Figure 1: Paired bootstrap resampling graphs using BLEU differences between *DSTF-EN* and *MML* (left graph) and using F-measure differences (right graph)

System	BS-strong	BS-IN	MML	DSTF-EN
BLEU	34.96	32.44	34.62	35.40
METEOR	35.56	34.51	35.42	35.54

Table 4: averaged BLEU scores for Spanish-English

According to the BLEU scores, our method outperformed both baselines and gained almost 1 BLEU point over *MML*. The strong baseline is very competitive with both data selection methods. This can easily be explained, since the system relies on the same interpolated language model as *DSTF-EN* and *MML*. There is a 3 BLEU points difference between our results and the baseline trained only the in-domain data and almost half a point BLEU score difference between the strong baseline and our method. With respect to the METEOR scores, our method again outperforms the state-of-the-art approach.

In order to determine whether our method (*DSTF-EN*) outperforms the state-of-the-art method (*MML*) from a statistical point of view, we applied paired bootstrap resampling (Koehn, 2004). The MTCompar-Eval tool (Klejch et al., 2015; Sudarikov et al., 2016) was used for this purpose where the source, reference and one or more system translations are used in the analysis. For our analysis we selected the best translation of each system according to their BLEU scores[8].

Figure 1 depicts the paired bootstrap resampling BLEU graph (left side) and the F-measure graph (right side). The x-axis is represented by 1000 resamples of the test set and the y-axis represents the difference in BLEU (respectively F-measure) between *DSTF-EN* and *MML* for all resamples. The p-value from the first graph in Figure 1 reports that in 11 cases out of the 1000 resamples, the state-of-the-art method performed better in terms of BLEU than our method (marked with a small red area in the graph). A similar behaviour can be oserved in the right graph from Figure 1 where in 34 cases out of 1000, *MML* outperformed *DSTF-EN* in terms of F-measure. Therefore in 96.6% of the times our method wins over the state-of-the-art when using the F-measure and in 98.9% of the cases, our method is better than *MML* when evaluating with BLEU (large green areas in the graphs). We conclude that our method has a statistical significant performance in comparison with the state-of-the-art method when selecting the 10% of the general domain sentences that were most similar to the in-domain.

5 Conclusions and Future Work

We introduced an unsophisticated data selection method based on word frequencies which scores general domain corpora in half an hour (on average when considering all general corpora for five language pairs). Our method yields good results in the WMT task, as well as in comparison with a state-of-the-art method and a strong baseline (for Spanish-English). Further analysis and experiments will be carried out in future work to assess whether the improvement of our method over the state-of-the-art that we observed for Spanish-English is also statistically significant for other language pairs.

[8]We tuned three times and averaged the BLEU scores

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. In *O'Reilly Media Inc*.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF. In *International Workshop on Spoken Language Translation, IWSLT 2005, Pittsburgh, PA, USA*, pages 61–67.

M F. Porter. 2001. Snowball: A language for stemming algorithms. In *Retrieved March*, volume 1.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. In *Cambridge, MA: MIT Press*.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 176–181.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proceedings of EAMT*, pages 133–142.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-Gram-Based Author Profiles For Authorship Attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING 2003*.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval : Graphical Evaluation Interface for Machine Translation Development. In *The Prague Bulletin of Mathematical Linguistics, Number 104*, pages 63–74.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for N-gram language modeling. In *Proceedings ICASSP*, pages 181–184.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP*.

Philipp Koehn. 2010. An Experimental Management System. *Prague Bull. Math. Linguistics*, 94:87–96.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan 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. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 65–72.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurelie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 Biomedical Translation Shared Task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Holger Schwenk and Philipp Koehn. 2008. Large and diverse language models for statistical machine

translation. In *In Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNP*.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Interspeech*, volume 2002.

Roman Sudarikov, Martin Popel, Ondrej Bojar, Aljoscha Burchardt, and Ondrej Klejch. 2016. Using MT-ComparEval. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Krzysztof Wolk and Krzysztof Marasek. 2014. Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs. In *Procedia Technology, 18*, pages 126 – 132. Elsevier.