Automatic Threshold Detection for Data Selection in Machine Translation

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

Abstract

We present in this paper the participation of the University of Hamburg in the Biomedical Translation Task of the Second Conference on Machine Translation (WMT 2017). Our contribution lies in adopting a new direction for performing data selection for Machine Translation via Paragraph Vector and a Feed Forward Neural Network Classifier. Continuous distributed vector representations of the sentences are used as features for the binary classifier. Most approaches in data selection rely on scoring and ranking general domain sentences with respect to their similarity to the in-domain and setting a range of thresholds for selecting a percentage of them for training various MT systems. The novelty of our method consists in developing an automatic threshold detection paradigm for data selection which provides an efficient and simple way for selecting the most similar sentences to the in-domain. Encouraging results are obtained using this approach for seven language pairs and four data sets.

1 Introduction

Data selection for Machine Translation (MT) represents a standard domain adaptation technique with the aim of tackling the problem of selecting from various general domain data the sentences that are most similar to sentences from the in-domain. Irrespective of having available vast amounts or small amounts of in-domain data, one of the advantages of data selection consists in providing more in-domain data selected from large amounts of general domain data. Two difficult tasks arise when performing data selection: what method to use for scoring the sentences from the general domain according to their similarity to the in-domain and how many of the scored sentences to keep for later use in training MT systems.

Standard state-of-the-art methods resolve the first difficulty by means of information retrieval, perplexity or edit distance methods. However, the second difficulty remains a challenge. There are no standard start-threshold and increment-threshold defined in the community. Axelrod et al. (2011), for example, uses the top N = {35k, 70k, 150k} sentence pairs from the scored general domain data, while Biçici and Yuret (2011) increasingly select N ∈ {100, 200, 500, 1000, 2000, 3000, 5000, 10000} instances for each test sentence for training and Kirchhoff and Bilmes (2014) select subsets of 10%, 20%, 30% and 40% of the data.

We present a time and resource efficient method of performing data selection using Paragraph Vector (Le and Mikolov, 2014) for representing the sentences and a Feed Forward Neural Network Classifier for determining which general domain sentences should be considered similar to the in-domain. The paragraph vectors and the binary classifiers are trained using standard parameters and have a great advantage of dropping the need to experiment with different sentence selection thresholds. Therefore, we call our method automatic threshold detection for data selection (ATD).

The method has been applied in the Biomedical translation task of the Second Conference on Machine Translation (WMT) 2017 (Yepes et al., 2017). The in-domain corpora were made available by the competition and the general domain corpora we have chosen to select data from are the Wikipedia corpora (Wolk and Marasek, 2014) and the Commoncrawl corpora[1]. Experiments

[1]http://commoncrawl.org/

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 483–488
Copenhagen, Denmark, September 711, 2017.

were performed on the language pairs English-French, English-Spanish, English-Portuguese and English-German (both directions for all language pairs except for English-German as the competition did not require German-English translations). Good results have been obtained for all language pairs.

The paper is structured as follows: related work is presented in Section 2, then the data, tools and data selection method are described in Section 3. Section 4 contains the experimental results and the last section presents conclusions and suggestions for future work.

2 Related work

Given a large pool of general domain data and a small amount of in-domain data, selecting the sentences from the general domain that are most similar to the in-domain is referred in literature as data selection. The work-flow of performing data selection includes developing a metric or function that scores general domain sentences according to their relevance to the in-domain and experimenting with various ratios of top ranked sentences in order to obtain the best result in terms of one or more MT evaluation metrics.

The approaches most commonly adopted in the literature are based on information retrieval (Hildebrand et al. (2005); Tamchyna et al. (2012)), on perplexity (Moore and Lewis (2010); Axelrod et al. (2011)), or on edit distance similarity (Wang et al., 2013).

Recently, a new direction has gained interest by making use of Word or Paragraph Vectors (embeddings). Chen and Huang (2016) use word embeddings along with in-domain selected sentences as positive samples and randomly selected sentences from the general domain as negative samples in training convolutional networks that yield good results. Also, Duma and Menzel (2016) developed a new scoring method using Paragraph Vectors with positive results.

In this paper, we apply Paragraph Vectors for training FFNN classifiers that categorize the general domain sentences as being in-domain or out-of-domain. One of the most challenging tasks in data selection consists in finding the optimal threshold (how many of the scored sentences to select). It is a time-consuming process in which several experiments need to be performed, usually aiming to obtain the best BLEU score. Moreover, there is no general consensus in the community regarding the increment ratio. We contribute to the state-of-the-art with a method that overcomes this challenge by means of a binary classifier: the problem of data selection is simplified by reducing the task of scoring and experimenting with different thresholds to a binary decision (keep/ discard a general domain sentence).

3 Experiments

This section describes the corpora and tools used, as well as the automatic threshold detection method we propose.

3.1 Data and tools

All SMT models were developed using the Moses phrase-based MT toolkit (Koehn et al., 2007) and the Experiment Management System (Koehn, 2010). The preprocessing of the data consisted in tokenization, cleaning (6-80), lowercasing and normalizing punctuation. The tuning and the test sets were provided by WMT 2016 (Bojar et al., 2016) and WMT 2017.

The SRILM toolkit (Stolcke, 2002) and Kneser-Ney discounting (Kneser and Ney, 1995) were used to estimate 5-gram language models (LM). All the trained SMT systems use a strong LM built by interpolating a LM for the in-domain and a LM for the general domain with weights that are tuned to minimize the perplexity on the tuning set (Schwenk and Koehn, 2008).

For word alignment we used GIZA++ (Och and Ney, 2003) with the default *grow-diag-final-and* alignment symmetrization method. Tuning of the SMT systems was performed with MERT (Och, 2003).

Commoncrawl and Wikipedia were used as general domains for all language pairs except for EN↔PT where no Commoncrawl data was provided by WMT. As for the in-domain corpora, EMEA (Tiedemann, 2012) was used for all language pairs and Muchmore, ECDC, Pattr and Pubmed (all from UFAL Medical Corpus[2]) for those language pairs where data was available. We also made use of the training data provided by the previous Biomedical task from 2016. The corpora corresponding to the general domain was concatenated into a single data source and the same procedure was applied for the in-domain corpora. The

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

size of the corpora is presented in the following table (since the bilingual corpora remain the same for both cases of translating $Language1$ to $Language2$ and vice-versa, we mention only one direction in the table):

Track / Corpora	EN-DE	EN-FR	EN-ES	EN-PT
Commoncrawl	2.4M	3.2M	1.8M	-
Wikipedia	2.4M	818K	1.8M	1.6M
EMEA	1.1M	1.09M	1.09M	1.08M
Muchmore	29K	-	-	-
ECDC	2547	2665	2357	-
Pattr	1.8M	-	-	-
Scielo-gma 2016	-	18K	175K	613K
Pubmed	-	-	285K	74K

Table 1: Corpora used for ATD

3.2 Automatic Threshold Detection for Data Selection

The data selection method we used for the WMT Biomedical task is described in this section with a special focus on Paragraph Vector and the FFNN classifier employed in developing the automatic threshold detection.

Paragraph Vector

Sentences were represented using Paragraph Vectors (Le and Mikolov, 2014) which give a continuous distributed vector representation of the input. Paragraph Vector is an extension of word embeddings (Mikolov et al., 2013) to phrases or sentences. Given a sentence, Paragraph Vector learns its representation by mapping context words and a paragraph identifier to the word to be predicted. The paragraph token acts like a memory of the topic of the sentence (Le and Mikolov, 2014). While the word vectors are shared between all paragraphs, the paragraph vector is shared among all the contexts generated from the same sentence.

We used the *gensim* toolkit[3] (Řehůřek and Sojka, 2010) that implements Doc2Vec (Paragraph Vectors). We present results using a Doc2Vec model trained with PV-DBOW[4] applying the default parameters of size 200 for the vectors and window of 10 (the maximum distance between the predicted word and context words used for prediction within a document).

[3]https://radimrehurek.com/gensim/models/doc2vec.html
[4]Distributed Bag of Words

Feed-forward Neural Network Classifier

The Feed-Forward Neural Network uses a supervised learning algorithm that receives as input the Paragraph Vectors for the labeled sentences. The feed-forward neural network classifier was trained using the python library *sknn*[5]. We report here results obtained using a fully connected *Tanh* layer of 200 units with dropout *p=0.5* and a *Softmax* output layer. The optimal dropout value was selected in accordance with the findings from Srivastava et al. (2014).

We experimented with both the source and the target language, in order to determine the best use of classified data given our settings.

For each of the language pairs we trained classifiers on ≈200K sentences with an equal number of positive and negative samples. The positive samples were randomly selected from the in-domain data and the negative samples were randomly selected from the general domain data.

4 Experimental results

We report in this section the BLEU (Papineni et al., 2002) scores obtained by our submissions, as well as the classifiers accuracy. For each language pair and for each test set provided by the Biomedical task, we submitted three runs as follows:

- the selected sentences with the classifier trained on the source language data (run 1)
- the selected sentences with the classifier trained on the target language data (run 2)
- the union (without duplicates) of the selected sentences proposed by the two classifiers (run 3)

Intrinsic evaluation of the proposed data selection technique was performed by computing the classifier accuracy. Following the recommendations from (Kohavi, 1995), we employ the stratified cross-validation method with ten folds. The accuracy values were computed using scikit-learn (Pedregosa et al., 2011). The following table presents the FFNN classifier mean accuracy and standard deviation for each of the language pairs. The low values of standard deviation for all classifiers indicate the consistency of our proposed method.

[5]http://scikit-neuralnetwork.readthedocs.io/en/latest/

Language pair	$FFNN_{source}$	$FFNN_{target}$
EN-DE	0.9715 ± 0.00085	0.9716 ± 0.00082
EN-ES	0.9403 ± 0.00221	0.9408 ± 0.00315
EN-FR	0.9585 ± 0.00364	0.9626 ± 0.00245
EN-PT	0.9596 ± 0.00197	0.9644 ± 0.00213

Table 2: Classifier accuracy (%): mean and standard deviation

This year four datasets were used in the evaluation: Scielo, EDP, Cochrane and NHS belonging to scientific publications or health information texts. The format of the datasets differed as Scielo and the EDP datasets follow the BioC format and Cochrane and NHS follow the format of the UFAL Corpus (sgm). Table 3 depicts the size of the datasets.

Language pair	Scielo	EDP	Cochrane	NHS
EN-DE	-	-	467	1044
EN-ES	1120	-	467	1044
ES-EN	1135	-	-	-
EN-FR	-	784	467	1044
FR-EN	-	662	-	-
EN-PT	1897	-	-	-
PT-EN	1825	-	-	-

Table 3: Size of the test sets

The results of our submissions are presented with respect to different datasets. Table 5 depicts all the BLEU scores of our submissions. For the Scielo dataset, our team was the only one that submitted runs. The organisers provided baselines for all language pairs and our best run improves with almost 9 BLEU points over the baseline for EN-PT and EN-ES, and almost 7 BLEU point over the baseline for PT-EN and ES-EN. There were small differences between the results of the three runs which suggests that either method could be used for gaining positive results.

For the EDP dataset (FR-EN and EN-FR) there were eight submissions and our best run for EN-FR had a gain of around 10 BLEU points over the baseline, as for FR-EN a gain of around 6 BLEU points. Considering our runs, there is 1 BLEU point difference between run 2 and run 3 for FR-EN and 0.5 difference between run 3 and run 2 for EN-FR. This indicates that the union method provides the best results.

On the Cochrane and NHS datasets our team was the only one that submitted for EN-ES obtaining high BLEU scores (48.99, 48.45 and 48.70 for Cochrane and 40.97, 41.20 and 41.22 for NHS). The differences between the runs are again very small. For EN-FR there were two teams participating. In our runs the union method gave better results for both datasets. For EN-DE there were six teams participating and the differences between our runs are again small.

In the general ranking among all participating teams, our team ranked first for EN-FR for the Cochrane and NHS datasets, second on FR-EN and third on EN-FR for the EDP datasets, last place on EN-DE for the Cochrane and NHS datasets, and was the only team submitting for Scielo (PT-EN, EN-PT, ES-EN, EN-ES) as well as for Cochrane and NHS (EN-ES).

Lavie (2010) points out that BLEU scores above 30 reflect understandable translations, while scores over 50 are considered good and fluent translations. Within 36 submitted runs by our team, 24 runs have BLEU scores between $\approx$32 and $\approx$49 (for six language pairs). Therefore, we conclude that the method presented obtains generally good translation results on a variety of language pairs.

Another important result consists in the fact that small amounts of general domain data were selected using ATD ranging from 3.1% up to 9.35%. This represents a promising direction for applying this method on much larger general domain corpora where selecting small amounts of data matters even more. The union of the selected sentences with the classifiers trained on the source and target languages ranges from 5.6% up to 12.1%.

The following table presents the amount of general data selected using ATD for the three runs along with the percentage of general domain data that it represents:

Language pair	# selected src. sent.	# selected trg. sent.	Union
EN-DE	148K (3.1%)	188K (4.0%)	263K (5.6%)
EN-ES	327K (9.35%)	257K (7.36%)	425K (12.1%)
EN-FR	223K (5.6%)	225K (5.7%)	345K (8.7%)
EN-PT	78K (4.7%)	89K (5.3%)	123K (7.4%)

Table 4: Number of selected sentences and percentage of General domain

The average duration for training the Doc2Vec models was $\approx$ 2.5 hours and the average duration for ten fold cross-validation was $\approx$ 12 minutes[6], which represents an advantage in terms of time consumption since afterwards only one MT system needs to be trained.

[6]on a 2 Ten Core Intel Xeon processor/ 128 GB of RAM machine

Language pair	EN-DE		EN-ES			ES-EN	EN-FR			FR-EN	EN-PT	PT-EN
Test set	Cochrane	NHS	Scielo	Cochrane	NHS	Scielo	EDP	Cochrane	NHS	EDP	Scielo	Scielo
run 1	22.03	18.71	36.08	**48.99**	40.97	37.14	22.43	32.46	31.79	22.64	39.14	43.84
run 2	22.37	**19.80**	35.93	48.45	41.20	37.47	22.25	32.59	31.89	22.37	**39.38**	**43.93**
run 3	**22.63**	19.66	**36.23**	48.70	**41.22**	**37.49**	**22.79**	**33.16**	**33.36**	**23.41**	39.21	43.88

Table 5: WMT results in terms of BLEU

5 Conclusions and Future Work

We presented the University of Hamburg participation to the WMT Biomedical task. The main contribution of our work consists in developing an automatic threshold detection method for data selection which yields good results for seven language pairs and four data sets. It requires little time for obtaining the general domain sentences that are considered most similar to the in-domain.

For six of the seven language pairs, the BLEU scores that our method obtained are in the range between 32 and 49. Generally, the best results among our three runs is obtained using the union approach, but with small differences among the other runs suggesting that there is no clear preference for one of the approaches.

Since we evaluated our approach only with respect to the WMT task, we intend to further apply it to other in-domains and language pairs, as well as, to compare it directly with standard state-of-the-art methods.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '11, pages 355–362. http://dl.acm.org/citation.cfm?id=2145432.2145474.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '11, pages 272–283. http://dl.acm.org/citation.cfm?id=2132960.2132996.

Ondřej 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*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301.

Boxing Chen and Fei Huang. 2016. Semi-supervised convolutional networks for translation adaptation with tiny amount of in-domain data. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. pages 314–323. http://aclweb.org/anthology/K/K16/K16-1031.pdf.

Mirela-Stefania Duma and Wolfgang Menzel. 2016. Data selection for IT texts using paragraph vector. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*. pages 428–434. http://aclweb.org/anthology/W/W16/W16-2331.pdf.

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.

Katrin Kirchhoff and Jeff A. Bilmes. 2014. Submodularity for data selection in machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 131–141. http://aclweb.org/anthology/D/D14/D14-1014.pdf.

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

Philipp Koehn. 2010. An experimental management system. *Prague Bull. Math. Linguistics* 94:87–96. http://dblp.uni-trier.de/db/journals/pbml/pbml94.html.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'95, pages 1137–1143. http://dl.acm.org/citation.cfm?id=1643031.1643047.

Alon Lavie. 2010. Evaluating the output of machine translation systems. In *AMTA*. Denver, Colorado, USA.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. pages 1188–1196.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*. pages 3111–3119.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACLShort '10, pages 220–224. http://dl.acm.org/citation.cfm?id=1858842.1858883.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '03, pages 160–167. https://doi.org/10.3115/1075096.1075117.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

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*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15(1):1929–1958. http://dl.acm.org/citation.cfm?id=2627435.2670313.

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

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2012. Selecting data for english-to-czech machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '12, pages 374–381. http://dl.acm.org/citation.cfm?id=2393015.2393068.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Junwen Xing, Yi Lu, and Isabel Trancoso. 2013. Edit distance: A new data selection criterion for domain adaptation in SMT. In *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*. pages 727–732. http://aclweb.org/anthology/R/R13/R13-1094.pdf.

Krzysztof Wolk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. In *Procedia Technology, 18*. Elsevier, pages 126 – 132.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods on Natural Language Processing (EMNLP)*. Copenhagen, Denmark.