UTFPR at WMT 2018: Minimalistic Supervised Corpora Filtering for Machine Translation

Gustavo H. Paetzold
Federal University of Technology - Paraná / Brazil
ghpaetzold@utfpr.edu.br

Abstract

We present the UTFPR systems at the WMT 2018 parallel corpus filtering task. Our supervised approach discerns between good and bad translations by training classic binary classification models over an artificially produced binary classification dataset derived from a high-quality translation set, and a minimalistic set of 6 semantic distance features that rely only on easy-to-gather resources. We rank translations by their probability for the "good" label. Our results show that logistic regression pairs best with our approach, yielding more consistent results throughout the different settings evaluated.

1 Introduction

It is no secret that Machine Translation (MT) systems have a wide array of applications, which range from translating news to multiple languages in order to more widely spread useful information, to producing translated transcriptions of real-time audio so that people from different places can communicate more easily.

MT systems have evolved considerably throughout recent years due mainly to the widespread adoption of neural machine translation (NMT) approaches. Attention-based encoder-decoders (Bahdanau et al., 2014) and neural semantic encoders (Munkhdalai and Yu, 2016) are just some examples of recurrent neural network architectures that have achieved great success in this task.

But regardless of how much MT approaches have evolved from a modelling standpoint, both modern and legacy approaches learn from the same type of information: parallel data containing hand-crafted translations. This data usually takes the form of millions (sometimes billions) of parallel original-to-translated sentences, and are often extracted from translated versions of documents, such as news articles (Bojar et al., 2017), and subtitles (Lison and Tiedemann, 2016).

Despite being hand-crafted, sometimes these datasets contain a lot of spurious translation examples that would not necessarily teach anything useful to an MT model, potentially compromising its performance. Consequently, it is important to filter these datasets in order to maximise the model's performance. Tiedemann (2012) and Lison et al. (2018) effectively filter large parallel corpora extracted from subtitles by using unsupervised metrics that combine features such as translation probabilities, language model probabilities, etc. In this contribution, we attempt to elaborate on the ideas of Tiedemann (2012) and Lison et al. (2018) by using such features as input to supervised machine learning models.

In what follows, we present the UTFPR systems for the WMT 2018 parallel corpus filtering task: A minimalistic approach that aims at combining easy-to-harvest features with classic supervised binary classification models to create efficient translation filters.

2 Task Description

The WMT 2018 parallel corpus filtering task is a very simple one: given a large dataset containing many automatically harvested translations, rank them according to their quality i.e. how useful one can expect them to be to an MT system.

The dataset provided contains around 1 billion words from English-to-German translations gath-

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 923–927
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64110

ered as part of the Paracrawl project (Buck and Koehn, 2016). The translations were of mixed domain, and among them are many spurious ones, such as misaligned translations, incomplete translations, translations with non-English and/or non-German sentences, etc. Participants were allowed to use the parallel corpora[1] from the WMT 2018 MT shared task to train their systems, if they wished to do so.

Participants were tasked with creating systems that assign a quality score to each translation in the dataset. To evaluate the systems, the organizers subsampled the dataset by choosing the N highest quality translations, training MT systems with them, then using traditional MT evaluation metrics to measure their performance. More details on the MT systems and evaluation metrics used are provided in Section 4.

3 Approach

In order to rank translations according to their quality, we've conceived a minimalistic supervised binary classification approach that relies on features that are easy to produce, and can hence be calculated even for resource-limited languages. The pipeline of our approach is illustrated in Figure 1.

First, we create a binary classification dataset using a set of high-quality English-German translations. The goal of this step is to create a very contrasting set of instances that greatly differed in terms of how coherent the source in English aligned with its German target. We create our dataset through the following steps:

1. We split the dataset in two equally sized portions, which we will henceforth refer to as "positive" and "negative" halves.

2. We then keep the positive half as it is, and shuffle the German side of the translations in the negative half, consequently misaligning the source and target side of the translations.

3. Finally, we assign label 1 (good quality) to all instances in the positive half, and -1 to the ones in the negative half (bad quality).

With our dataset at hand, we then calculate 6 features for each instance:

- The cosine distance between the average embedding vector of all content words in the source and target sentences.

- The minimum, maximum, and average cosine distance between the word embeddings of all possible word pairs in the source and target sentences.

- The proportion of words in the English source that have at least one ground-truth translation in the German target according to a dictionary.

- The proportion of words in the German target that have at least one ground-truth translation in the English source according to a dictionary.

These features have the main goal of capturing the overall semantic distance between the source and target in different ways. Notice that, since we prioritised creating an efficient and extensible approach to this task, we refrained from trying to exploit other features that attempt to capture syntactic properties, which require for parsers, which are often scarce for resource-limited languages.

To calculate our cosine distance features, we use the pre-trained 300-dimension English-German bilingual embeddings made available by the MUSE project (Lample et al., 2017). These embeddings offer a common distributional feature space for both English and German, and allow for us to calculate the cosine distance between English and German words. For the translation precision features, we used the English-German ground truth dictionary also made available by the MUSE project. These dictionaries are derived in unsupervised fashion from the same learning process that originate the previously described embeddings. Both of these resources can be obtained with raw text, without the need for parallel corpora, which makes our features easily obtainable for the great majority of languages. We treat as content words any words that are not featured in a list of stop words.

After feature calculation, we train a binary classification model over our dataset. At test time, we produce quality scores for unseen instances by calculating the same 6 features, passing them through our model, then extracting the probability of the positive class (label 1). To create a set of filtered translations, we rank the translations according to

[1] http://statmt.org/wmt18/translation-task.html

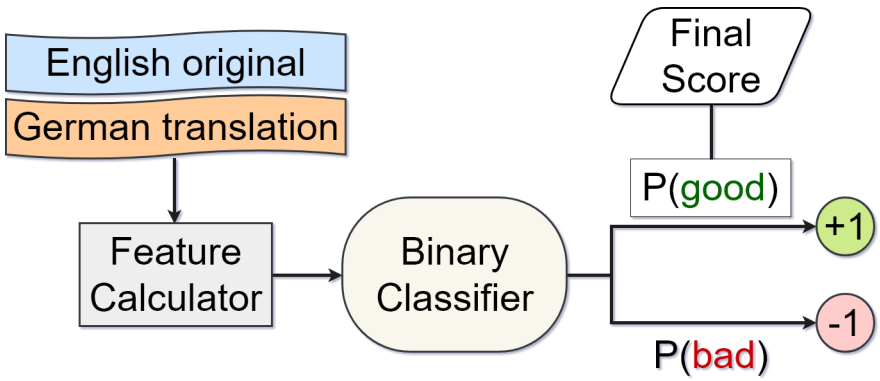

Figure 1: Architecture of the UTFPR systems

their positive class probabilities and choose the ones with highest scores. We name our approach UTFPR in reference to the university sponsoring this contribution.

4 Experimental Setup

As mentioned in Section 2, we submit our results to the parallel corpus filtering shared task of WMT 2018, of which the test set contains roughly one billion unfiltered parallel English-German translations. To train our supervised model, we use the Europarl v7 parallel corpus (Koehn, 2005), which contains $1,920,209$ translations.

For learning, we experiment with three classification models: Logistic Regression (UTFPR-LR), Decision Trees (UTFPR-DT), and Random Forests (UTFPR-RF). We chose them because they use a varying array of learning methods, and can be trained efficiently even when presented with hundreds of millions of input instances.

To evaluate our approach, the shared task organizers first created two sub-sampled sets of parallel translations containing the 10 million and 100 million highest scoring translations in the test set. They then used these sets to train both statistical (SMT) and neural MT (NMT) models using the Moses (Koehn et al., 2007) and Marian (Junczys-Dowmunt et al., 2018) toolkits, and evaluated the models according to BLEU-c (Koehn, 2011) over a combination of the newstest 2018[2], iwslt 2017[3], Acquis[4], EMEA[5], Global Voices[6], and KDE[7] datasets.

[2]http://statmt.org/wmt18/translation-task.html
[3]https://sites.google.com/site/iwsltevaluation2017
[4]https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[5]http://opus.nlpl.eu/EMEA.php
[6]http://opus.nlpl.eu/GlobalVoices.php
[7]http://opus.nlpl.eu/KDEdoc.php

5 Results

We compare our approach to the 5 systems from the WMT 2018 parallel corpus filtering task with the highest and lowest average BLEU-c scores. The results illustrated in Table 1 reveal that, although our models do not fair very well against more sophisticated strategies, they do perform more consistently than other strategies of similar performance across all the settings evaluated; one can observe that the main reason why our logistic regressor outperforms the bottom five shared task systems is because it achieves similar BLEU-c scores in all settings, while the bottom five achieve unusually low BLEU-c scores in some settings (particularly 10M sentences for NMT). However, this is not necessarily a strong point of our approach, since one would expect to achieve significantly higher scores in settings where the MT systems are being fed more sentences, specially in the case of NMT. This suggests that our models may be prone to choosing redundant/repetitive content.

It can also be noted that, overall, the logistic regression model performs much better than both our decision trees and random forests, specially for NMT, where the difference between them reaches upwards of 16.08 BLEU-c points. Inspecting the highest scores produced by these models, we found that our logistic regressor and the tree-based models prioritise much different translations. Both our decision tree and random forest assign higher scores to very short translation pairs averaging 15 tokens in length on either side, while our logistic regressor prioritises much longer ones, averaging 40 tokens in length on either side. We noticed that, although the shorter translation pairs prioritised by our tree-based models often feature a slimmer array of translation errors, they seem much less use-

	SMT		NMT		
	10M	**100M**	**10M**	**100M**	**Average**
Microsoft	24.45	26.50	28.62	32.06	27.91
RWTH	24.58	26.21	28.01	31.29	27.52
Alibaba	24.11	26.44	27.60	31.93	27.52
Alibaba-Div	24.11	26.42	27.60	31.92	27.51
NRC	23.89	26.40	27.41	31.88	27.39
UTFPR-LR	**20.81**	**22.35**	**21.75**	**22.23**	**21.79**
UTFPR-DT	17.55	20.67	11.44	11.88	15.38
UTFPR-RF	13.22	16.96	6.57	6.15	10.72
AFRL-Small	21.93	22.89	13.49	21.05	19.84
DCU-System 4	15.67	21.19	6.27	18.60	15.43
DCU-System 3	15.26	21.09	5.01	18.39	14.94
DCU-System 2	12.86	18.57	3.42	8.61	10.86
DCU-System 1	6.56	13.22	3.34	4.78	6.98

Table 1: Parallel corpus filtering results with respect to the average BLEU-c scores obtained over the datasets described in Section 4. The first and last five lines feature, respectively, the five systems that achieved the highest and lowest average BLEU-c scores in the task. Boldface numbers highlight the highest BLEU-c scores achieved among the UTFPR systems.

ful to an MT system. Most of them are translations of dates, article titles, ads, and list items, which we expect would offer little to no insight on how to translate longer, more elaborate sentences. In contrast, the longer translations prioritised by our logistic regressor feature more meaningful, complex sentences, which is most likely why they make for better input to MT models.

6 Conclusions

In this contribution, we presented the UTFPR systems submitted to the WMT 2018 parallel corpus filtering task. Our supervised systems discern between good and bad translations using classic binary classification models, and use as input a minimalistic set of 6 features that aim to capture the semantic distance between original and translated sentences without relying neither on syntactic information or scarce resources and tools.

We found that our approach performs best when employing logistic regression. Overall, our best performing system places 41th, when considering the BLEU-c average of all outcomes evaluated. In the future, we aim to evaluate the effectiveness of applying more elaborate dataset creation methods for training that produce more types of errors, employing more sophisticated neural models for the task, and incorporating cost-effective syntactic clues into the feature set.

7 Acknowledgments

We would like to thank the Federal University of Technology - Paraná for supporting this contribution.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the 2nd Conference on Machine Translation*, pages 169–214.

Buck, Christian and Philipp Koehn. 2016. Findings of the wmt 2016 bilingual document alignment shared task. In *Proceedings of the 1st Conference on Machine Translation*, pages 554–563.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of the 56th annual meeting of the ACL*, pages 116–121.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source

toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL*, pages 177–180.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, Philipp. 2011. What is a better translation? reflections on six years of running evaluation campaigns. *Tralogy 2011*, page 9.

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lison, Pierre and Jrg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th LREC*.

Lison, Pierre, Jrg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th LREC*.

Munkhdalai, Tsendsuren and Hong Yu. 2016. Neural semantic encoders. *CoRR*, abs/1607.04315.

Tiedemann, Jrg. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th LREC*.