Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation

Marine Carpuat and **Yogarshi Vyas** and **Xing Niu**
Department of Computer Science
University of Maryland
College Park, MD
{marine,yogarshi,xingniu}@cs.umd.edu

Abstract

Parallel corpora are often not as parallel as one might assume: non-literal translations and noisy translations abound, even in curated corpora routinely used for training and evaluation. We use a cross-lingual textual entailment system to distinguish sentence pairs that are parallel in meaning from those that are not, and show that filtering out divergent examples from training improves translation quality.

1 Introduction

Parallel sentence pairs provide examples of translation equivalence to train Machine Translation (MT) and cross-lingual Natural Language Processing. However, despite what the term "parallel" implies, the source and target language often do not convey the exact same meaning. This is a surprisingly common phenomenon, not only in noisy corpora automatically extracted from comparable collections, but also in parallel training and test corpora, as can be seen in Table 1.

This issue has mostly been ignored in machine translation, where parallel sentences are assumed to be translations of each other, and translations are assumed to have the same meaning. Prior work on characterizing parallel sentences for MT has focused on data selection and weighting for domain adaptation (Foster and Kuhn, 2007; Axelrod et al., 2011, among others), and on assessing the relevance of parallel sentences by comparison with a corpus of interest. In contrast, we focus on detecting an intrinsic property of parallel sentence pairs. Divergent sentence pairs have been viewed as noise both in comparable and non-parallel corpora (Fung and Cheung, 2004; Munteanu and Marcu, 2005; AbduI-Rauf and Schwenk, 2009; Smith et al., 2010; Riesa and Marcu, 2012) and in parallel corpora (Okita et al., 2009; Jiang et al., 2010; Denkowski et al., 2012).In contrast, we hypothesize that the translation process inherently introduces divergences that affect meaning, and that semantically divergent examples should be expected in all parallel corpora.

Divergent segments in OpenSubtitles	
en	someone wanted to cook bratwurst.
fr	vous vouliez des saucisses grillées.
gl	you wanted some grilled sausages.
en	i don't know what i'm gonna do.
fr	j'en sais rien.
gl	i don't know.
en	- has the sake chilled? - no, it's fine.
fr	- c'est assez chaud?
gl	- it is hot enough?
en	you help me with zander and i helped you with joe.
fr	tu m'as aidée avec zander, je t'ai aidée avec joe.
gl	you helped me with zander, i helped you with joe.
Divergent segments in newstest2012	
en	i know they did.
fr	je le sais.
gl	i know it.
en	the female employee suffered from shock.
fr	les victimes ont survécu leur peur.
gl	the victims have survived their fear.

Table 1: Parallel segments are not always semantically equivalent, as can be seen in these examples (English sentence (en), French sentence (fr) and its gloss (gl)) drawn from a random sample of OpenSubtitles and of the newstest2012 test set (Bojar et al., 2016).

We show that semantically divergent examples significantly impact the learning curves and translation quality of neural machine translation systems. We repurpose the task of cross-lingual textual entailment (Mehdad et al., 2010) to automatically identify and filter divergent parallel sentence pairs from the OpenSubtitles corpus (Lison and Tiedemann, 2016). This approach outperforms

Proceedings of the First Workshop on Neural Machine Translation, pages 69–79,
Vancouver, Canada, August 4, 2017.

other data selection criterion, and even a system trained on twice as much data for two test genres.

2 Non-Divergence as a Data Selection Criterion

2.1 Motivation

We conjecture that training sequence-to-sequence models with attention for neural machine translation (Bahdanau et al., 2015; Sennrich et al., 2017) is more sensitive to divergent parallel examples than traditional phrase-based systems. Phrase-based systems are remarkably robust to noise in parallel segments: Goutte et al. (2012) showed that when introducing noise by permuting the target side of parallel pairs, as many as 30% of training examples had to be noisy to hurt BLEU score significantly. However, such artificial noise does not capture naturally occurring divergences, which are likely to be more subtle. Syntax-based systems have been shown to be sensitive to divergences when they generate word-alignment errors: for instance, using syntax to eliminate word alignment links that violate syntactic correspondences yields better string-to-tree transducer rules, and better translation quality (Fossum et al., 2008).

In contrast, there is evidence that deep neural networks are sensitive to the nature and order of training examples in various related settings. On image classification benchmarks, Zhang et al. (2017) show that convolutional neural networks have the capacity to memorize versions of the training data corrupted in various ways, including random labelings of the original images, and random transformations of the input images. This suggests that neural models might attempt to memorize the idiosyncracies of divergent parallel segments, which might hurt generalization at test time. In machine translation, domain adaptation results (Durrani et al., 2016) show that neural models benefit from early training on the United Nations corpus before fine-tuning on in-domain data, while the UN corpus is generally considered to be too distant from any domain that is not UN to be useful when training e.g., phrase-based systems. Online training also motivates curriculum learning approaches: ordering examples from easier short sentences to harder long sentences has also been found advantageous for neural language modeling (Bengio et al., 2009).

Most directly related to this work, Chen et al. (2016) suggest that neural MT systems are more sensitive to sentence pair permutations than phrase-based systems (Goutte et al., 2012). They also show that a bilingual convolutional neural network trained to discriminate in-domain from out-of-domain sentence pairs effectively selects training data that is not only in domain but also less noisy. These results provide further evidence that the degree of parallelism in training examples has an impact in neural MT. Yet it remains to be seen to what extent semantic divergences – rather than noise – affect translation quality in general – and not only in domain adaptation settings.

2.2 Approach

In this paper, we seek to measure the impact of semantic divergence on translation quality when used as a data selection criterion: if our hypothesis holds, then training on non-divergent examples should yield better translation quality than training on the same number of examples selected using other criteria. Unlike in domain adaptation, semantic divergence is an intrinsic property of a parallel sentence pair, and is therefore independent of domains or specific testing conditions. As we will see, we treat the detection of the divergent examples as a classification problem. Training examples can be ranked based on the confidence of the classifier that the segment contains two sentences that are not equivalent in meaning, and use the resulting ranking to filter out examples.

In addition, data selection can help address practical concerns. Training neural machine translation systems on large scale parallel corpora has a prohibitive computational cost. For instance, the winning neural systems at the WMT evaluation required two weeks of training (Sennrich et al., 2016a). Automatically identifying the most useful training examples has the potential to reduce training time significantly.

3 Detecting Semantic Divergences in Parallel Segments

We aim to automatically detect whether the source and target side of a parallel example are semantically equivalent. Since parallel corpora are not readily annotated with semantic equivalence, we repurpose related cross-lingual semantic annotations and models for this task.

3.1 Model

We frame the task of detecting whether parallel sentences (e, f) are equivalent as a classification problem. We draw inspiration from related work on semantic textual similarity (Agirre et al., 2016), translation quality estimation (Hildebrand and Vogel, 2013), parallel sentence detection (Munteanu and Marcu, 2006) to design simple features that can be induced without supervision.

First, differences in sentence lengths are strong indicators of divergence in content between e and f. Accordingly, we use four length features: $|e|$, $|f|$, $\frac{|e|}{|f|}$, and $\frac{|f|}{|e|}$.

Second, we assume that the configuration of word alignment links between parallel sentences (e, f) is indicative of equivalence: if e and f have the same meaning, then they will be easier to align. Accordingly, we compute the following features for each of e and f:

- Ratio of aligned words
- Ratio of unaligned words
- Ratio of unaligned content words (defined as words that do not appear in a stopword list)
- Number of unaligned contiguous sequences
- Length of longest contiguous unaligned sequence
- Average length of aligned sequences
- Average length of unaligned sequences

3.2 Semantic Supervision

We use annotations of Cross-Lingual Textual Entailment (Mehdad et al., 2010). This task is framed as a four-way classification task. Given sentences e and f, the goal is to predict whether (1) e entails f, (2) f entails e, (3) e and f both entail each other, (4) there is no entailment relation between e and f. Negri and Mehdad (2010) showed that English training and test sets can be created by crowdsourcing, that are then translated to obtain cross-lingual datasets. Training and test data were made available at SemEval 2012 and 2013 (Negri et al., 2012, 2013). We hypothesize that examples detected as class (4) are the most divergent examples that are the least useful for training machine translation systems. While the 4-way classification task is more complex than our end goal of detecting divergent examples, we found that the 4-way classifier detects divergent examples from class (4) better than binary classifiers trained on various partitions of the 4-way training data.

Other relevant semantic annotations of bilingual corpora include cross-lingual semantic textual similarity (Agirre et al., 2016) and machine translation quality estimation datasets (Specia et al., 2010). The latter is not a good fit as it annotates machine translation output. The former is a better match but only provides test examples.

4 Experimental Settings

4.1 Divergence Detection Model Settings

We use the cross-lingual textual entailment datasets released at SemEval (Negri et al., 2012, 2013). The 2012 dataset consists of 1000 sentences per language, with equal train and test splits, while the 2013 dataset consists of 1500 sentences per language, 500 of which have been marked as the test set. All datasets are balanced across the four entailment classes.

Word alignments for features are trained on the Europarl corpus and the News Commentary corpus [1], with a total of 2.2M sentence pairs. We use symmetrized IBM 4 alignments obtained via MGIZA++, and obtain alignments for the CLTE data as well as the OpenSubtitles data by transductive training.

The classifier is the linear SVM implementation from Scikit-Learn [2] with $C = 1.0$ and a one-vs-rest multi-class scheme for 4-way classification.

4.2 Machine Translation Task and System

We evaluate on the English-French Microsoft Spoken Language Translation task (Federmann and Lewis, 2016), which provides a translation scenario motivated by real world applications. Following prior work (Farajian et al., 2016; Lewis et al., 2016), training data is drawn from the OpenSubtitles corpus (Lison and Tiedemann, 2016). The subtitle genre is also appropriate as it presents many potential divergences due to genre-specific constraints (Tiedemann, 2007). In addition, the robustness of the models is evaluated by testing on a second domain, leveraging publicly available TED talks test data (Cettolo et al., 2012). French and English sides of the corpus are uncased and tokenized using Moses preprocessing tools, and segmented using byte pair encoding (Sennrich et al., 2016b).

[1] http://www.statmt.org/wmt15/training-parallel-nc-v10.tgz

[2] http://scikit-learn.org/stable/

Corpus	# Sentences	English		French	
		|vocab|	# tokens	|vocab|	# tokens
Training Sets (Open Subtitles)					
non-divergent	17M	63917	133.7M	79818	133.6M
random	17M	64640	146.9M	80222	139.1M
natural	17M	64495	147.3M	77646	137.4M
length < 10	22M	63643	133.0M	79264	127.2M
all	33.5M	66935	288.5M	82564	273.2M
Test Sets					
MSLT	5292	3739	45197	4389	49562
TED	1305	3987	25466	4481	27513

Table 2: Data statistics for training and test sets. At train time, selecting non-divergent sentences yields (1) a smaller vocabulary compared to datasets of the same size (2) richer examples than selection based on length only, with a more diverse vocabulary.

The neural machine translation system is the encoder-decoder with attention implemented in Nematus (Sennrich et al., 2017), with suggested parameters. We use a vocabulary size of 90000, dimensions of word embeddings and hidden units are 500 and 1024 respectively. Models are trained using Adadelta (Zeiler, 2012), with a learning rate of 0.0001, a batch size of 80, and reshuffling at each epoch. Dropout is enabled. We use the first 5th of the MSLT test set as a validation set, and save models every 30000 updates.

4.3 Experimental Conditions

We empirically evaluate the impact of divergence on translation quality by considering the following experimental conditions, which correspond to different training sets for the same neural MT model and training configuration:

- NON-DIVERGENT filtering out the most divergent half of the training data
- RANDOM randomly downsampling the training corpus to half its size
- NATURAL use the natural order of the corpus files to select the first half of the corpus
- LENGTH select examples of length shorter than 10 words (the average sent length in the corpus)
- ALL default condition which uses the entire training corpus.

Training data statistics and their coverage of the test set are summarized in Table 2. Data selection naturally reduces the vocabulary size available compared to using all the training data, by at most 10%. Selecting non-divergent sentences yields a smaller vocabulary compared to using the same number of parallel sentence pairs selected based on natural order or random sampling. At the same time, non-divergent examples are richer than those selected based on length alone, with a more diverse vocabulary.

Test data statistics shows the complementarity of the two test conditions considered: the MSLT task consists of shorter sentences similarly to all training settings, while the TED tasks consists of much longer segments.

5 Experiment Results

5.1 Preliminary Check on Divergence Detection

The supervised classifier based on simple features yields competitive performance with published Cross-Lingual Textual Entailment results. On the 2012 test set, it achieves an accuracy of 60.4% , outperforming the best published result of 57% (Jimenez et al., 2012). On the harder 2013 test set, it achieves 43.6%, approaching the best published result of 45.8% (Zhao et al., 2013).

As a sanity check, we annotate a small sample of 100 randomly selected examples from OpenSubtitles. A bilingual speaker was asked to evaluate whether parallel segments in the two languages have exactly the same meaning or not. Surprisingly as many as 37% of examples were found to diverge in meaning. The nature of the divergences vary, but can generally be explained by discourse and explicitation effects (see Table 1).

The classifier detects semantically divergent

sentence pairs with a precision of 62.5% and a recall of 13.5%. The low recall shows that there is room to improve divergence detection, including by enriching the model, exploring alternate sources of supervision and adapting to the domain of the parallel data classified. Nevertheless, given that the default MT set-up consists in using all divergent sentences (i.e. detecting divergent sentences with a precision and recall of 0%), the current model represents a significant improvement.

5.2 Impact on Translation Quality

The learning curves (Figure 1) show that better translation quality can be achieved faster using NON-DIVERGENT as a selection criteria, even when compared to models trained on more data. On the validation set, the NON-DIVERGENT model achieves the BLEU score of the ALL model with only 60% of the updates.

RANDOM data selection yields a curve that is close to that of the ALL model. Selecting the first half of the corpus (NATURAL) plateaus about 6 points lower than the best models. The stark difference in performance between RANDOM and NATURAL might be explained by the fact that the RANDOM training set contains a more diverse set of sentences, sampled from a broader range of movies than the NATURAL dataset. This is supported by the corpus statistics in Table 2 which show that the RANDOM training set has a larger vocabulary size than the NATURAL one, especially in French. Training only on short sentences ($<$ 10 words) does much worse as the resulting system produces short translations which trigger high BLEU brevity penalties.

Table 3 shows the translation quality of the systems considered on two test sets using ensemble decoding. Following Sennrich et al. (2016a), translations are obtained by decoding with an ensemble of the 3 best models saved during training. The NON-DIVERGENCE criterion yields the best BLEU scores on both test sets, and even outperforms the system trained on all data by +1.6 BLEU on the MSLT task and by +0.6 BLEU on the TED task. System relative rankings are overall consistent with the learning curve: the NON-DIVERGENT system is best, either the ALL or RANDOM system are in 2nd or 3rd place depending on the test set, and using the NATURAL order of the corpus does much worse. Training on short sentences hurts in both cases, but particularly on the TED task which consists of longer segments.

System	TED	MSLT
best mix (all data)	33.03	40.11
best mix (non-divergent)	34.23	41.74
+ best model (all data)	**34.57**	**42.13**

Table 5: Ensembles of systems (mix) trained on all data and non-divergent data yield modest improvements in BLEU

5.3 Impact of Longer Training

One might wonder whether the trends above still hold when training longer, since training is expected to take longer to converge with more examples to learn from. We therefore continue training for all promising models (i.e. all but the system trained on short sentences only). Figure 5.3 shows that learning curves for NON-DIVERGENT, RANDOM and ALL eventually converge. However, Table 4 show that, among systems trained on the same number of examples, NON-DIVERGENCE remains the best data selection criterion, and that it yields decoding results that continue to outperform ensembles of models trained on ALL.

5.4 Ensembles of Models from Multiple Training Conditions

Finally, we evaluate whether models trained on ALL and on the NON-DIVERGENT data are complementary by augmenting the best performing systems in Table 4, which are all ensembles of models trained on non-divergent data, with the best model trained on the entire training set. Table 5 shows that the mixed ensemble improves over the previous best result by +0.34 BLEU on the TED test set and +0.40 on the MSLT test set. It is unclear whether these modest gains are worth the additional training time needed to add the ALL system to the mix. However, it remains to be seen whether better model selection could yield further improvements.

6 Related Work

Translation Divergences Most prior work on translation divergences has focused on typological issues which reflect the fact that languages do not encode the same information in the same way. Dorr (1994) formalizes this problem by defining divergence categories (e.g., thematic, structural,

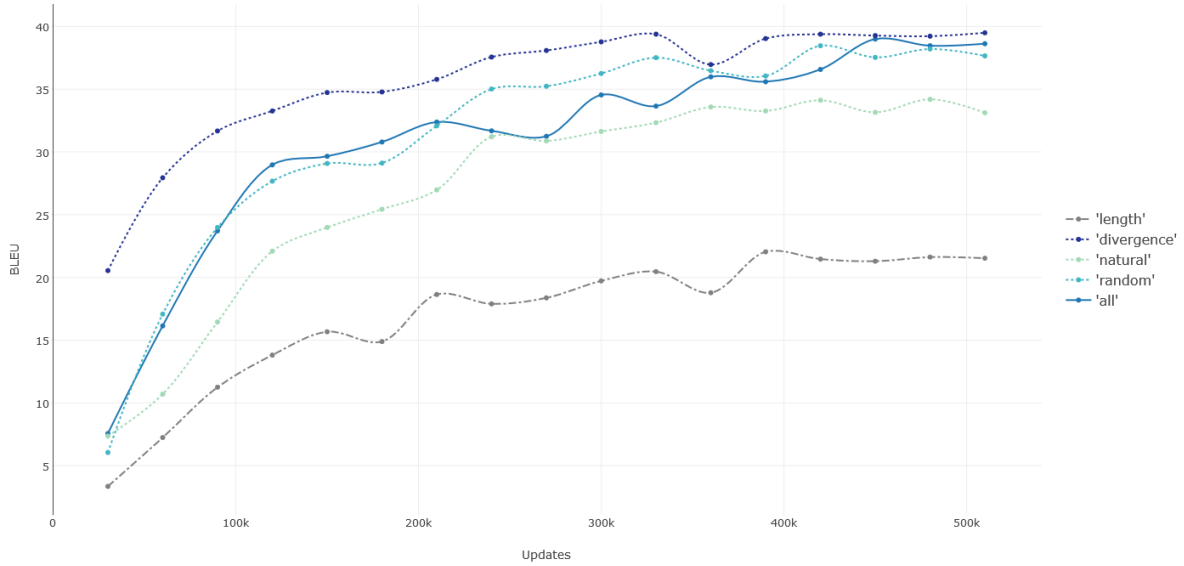

Figure 1: Learning curves on validation test set for all training configurations. Solid lines indicate a system trained on the entire training set, dotted lines use half of the training data with various selection criteria, and the dashed line indicates data selected by length. Training on the non-divergent half of the examples yields the top learning curve, even when compared to using all of the training data.

Selected Data	**Size (M)**	**TED**		**MSLT**	
		3-best	3-last	3-best	3-last
non-divergent	17	**32.47**	**32.90**	**40.27**	**40.37**
random	17	31.42	31.54	39.03	39.23
natural order	17	29.76	30.26	35.57	35.59
length < 10	22	8.25	8.25	22.55	22.55
all	33.5	31.88	32.12	38.70	38.60

Table 3: Impact of data selection criterion on TED and MSLT test sets translated by an ensemble of the 3 best models saved during training. Filtering out DIVERGENT examples yields the best translation quality.

Selected Data	**Size (M)**	**TED**		**MSLT**	
		3-best	3-last	3-best	3-last
non-divergent	17	**33.90**	**34.23**	**41.74**	**41.24**
random	17	33.03	33.51	39.64	39.64
natural order	17	31.94	32.29	37.47	36.94
all	33.5	33.03	33.03	40.11	40.11

Table 4: Impact of longer training time on BLEU scores for TED and MSLT test sets translated by ensembles of 3 models. Filtering out divergent examples still yields the best translation quality, outperforming other selection criteria as well as systems trained on all data.

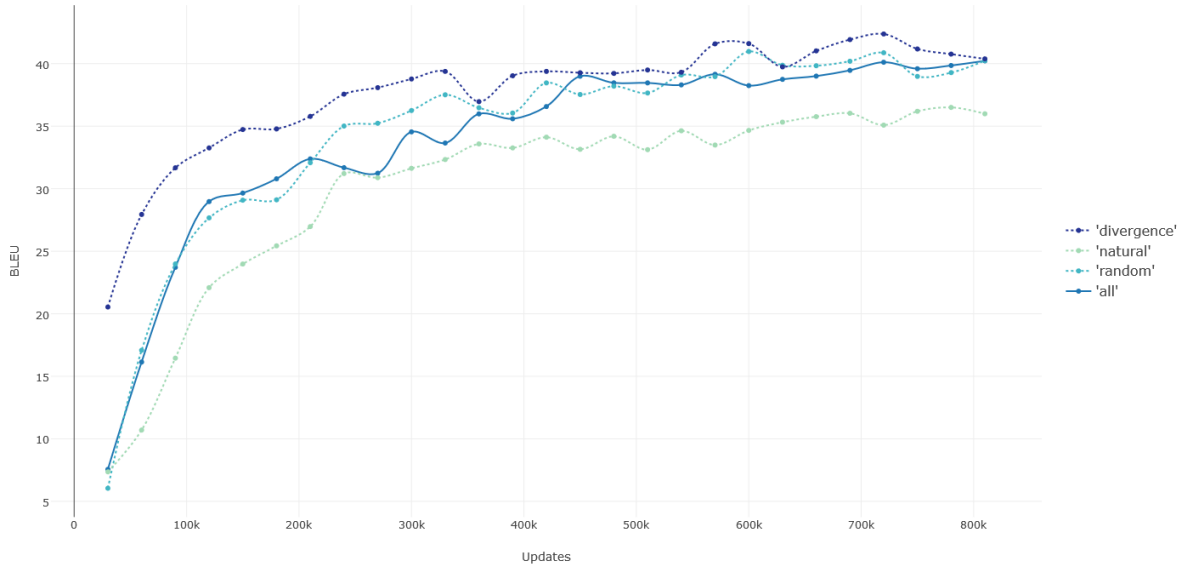

Figure 2: Longer learning curves on validation test set for most promising models. Training on NON-DIVERGENT examples remains the top system although RANDOM and ALL eventually converge to the same validation set performance.

categorical). Follow up work shows that divergences are not outliers but common phenomena in parallel corpora (Dorr et al., 2002; Habash and Dorr, 2002). Some of these divergences have been implicitly addressed by designing MT architectures informed by syntax and structure (Wu, 1997; Habash and Dorr, 2002; Chiang, 2007; Lavie, 2008, among others). In this work, we focused instead on semantic divergences which happen when the source and target sentences do not convey exactly the same meaning.

Modeling Cross-Lingual Semantic Divergences Prior work has addressed cross-lingual semantic textual similarity (Agirre et al., 2016), entailment (Negri and Mehdad, 2010; Negri et al., 2012, 2013), translation quality estimation (Specia et al., 2010, 2016). While the human judgments obtained for each task differ, all tasks take inputs of the same form (two segments in two different languages) and output a prediction that can be interpreted as indicating whether they are equivalent in meaning or not. Models share core intuitions, relying either on MT to turn the cross-lingual task into its monolingual equivalent, or on features derived from MT components such as translation dictionaries and word alignments.

Extracting Parallel Sentences from Non-Parallel Corpora Extracting parallel sentences or parallel fragments from non-parallel corpora differs from our work in several ways. The goal is to identify additional training examples to augment parallel corpora, rather than to identify the most useful examples in a parallel corpus (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2005; AbduI-Rauf and Schwenk, 2009; Smith et al., 2010; Riesa and Marcu, 2012). The non-parallel examples tend to be more extreme than in the parallel corpora considered in our work.

Data Cleaning This line of work aims to remove noise, e.g., from alignment errors, based on scores from word alignment or language models (Okita et al., 2009; Jiang et al., 2010; Denkowski et al., 2012; Matthews et al., 2014). Cleaning training data in high-resource settings (Denkowski et al., 2012) and tuning data in lower resource settings (Matthews et al., 2014) has been shown to improve hierarchical phrase-based systems.

Incorporating Word Alignments into Neural MT Since our data selection criterion relies on word alignments, one could view our approach as part of the family of models that seek to im-

prove neural machine translation using insights and models from word alignment and statistical machine translation models (Cohn et al., 2016; Mi et al., 2016). These approaches however focus on improving neural machine translation in low resource settings, while our aim was to identify a subset of examples in large training sets.

Applications beyond MT Detecting cross-lingual semantic divergences using entailment has been motivated by the need to synchronize content across languages in multilingual resources such as Wikipedia (Negri and Mehdad, 2010; Duh et al., 2013). It could also be useful to select better training examples for cross-lingual transfer learning of semantic models (Yarowsky et al., 2001; Ganchev and Das, 2013, among others).

7 Conclusion

We showed that neural machine translation is sensitive to semantically divergent parallel segments, as detected by a simple cross-lingual textual entailment system. When controlling for the number of training examples, filtering out divergent segments yields significantly better translation quality than using a random sample of examples, or short examples. Selecting non-divergent examples also improves translation quality compared to a system trained on twice as much data.

In future work, we will extend our empirical study to a broader range of tasks including more distant language pairs than English-French and a range of training domains in addition to subtitles. We will also evaluate whether our findings are impacted by the choice of optimizer, since it has been shown to have an impact on the initial performance and convergence of models on constant training data (Farajian et al., 2016). Furthermore, we will aim to answer two open questions raised by these promising results: can the cross-lingual entailment detector be replaced by a more direct approach for detecting divergence? And to what extent are alignment-based features useful when compared to neural models that might be closer to that of neural machine translation systems?

Acknowledgments

We thank the CLIP lab at the University of Maryland and the reviewers for their constructive feedback. This work was supported in part by research awards from Amazon and Google.

References

Sadaf AbduI-Rauf and Holger Schwenk. 2009. On the Use of Comparable Corpora to Improve SMT Performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Athens, Greece, EACL '09, pages 16–23.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 497–511.

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, Edinburgh, United Kingdom, EMNLP '11, pages 355–362.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, Montreal, Quebec, Canada, ICML '09, pages 41–48. https://doi.org/10.1145/1553374.1553380.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, and others. 2016. Findings of the 2016 conference on machine translation (WMT16). In *Proceedings of the First Conference on Machine Translation (WMT)*. volume 2, pages 131–198.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. volume 261, page 268.

Boxing Chen, Roland Kuhn, George Foster, Colin Cherry, and Fei Huang. 2016. Bilingual Methods for Adaptive Training Data Selection for Machine Translation. *Proceedings of the 12th Conference of the Association for Machine Translation in the Americas (AMTA)* page 93.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics* 33(2):201–228.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model. In *Proceedings of NAACL-HLT*. pages 876–885.

Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The CMU-avenue French-English Translation System. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '12, pages 261–266.

Bonnie J. Dorr. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computional Linguistics* 20(4):597–633.

Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*. Springer-Verlag, London, UK, UK, AMTA '02, pages 31–43.

Kevin Duh, Ching-Man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing Information Disparity in Multilingual Document Collections. *ACM Trans. Speech Lang. Process.* 10(1):1:1–1:28. https://doi.org/10.1145/2442076.2442077.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2016. QCRI Machine Translation Systems for IWSLT 16. *International Workshop on Spoken Language Translation (IWSLT)* .

M. Amin Farajian, Rajen Chatterjee, Costanza Conforti, Shahab Jalalvand, Vevake Balaraman, Mattia A. Di Gangi, Duygu Ataman, Marco Turchi, Matteo Negri, and Marcello Federico. 2016. FBK's Neural Machine Translation Systems for IWSLT 2016. In *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT), Seattle, WA*.

Christian Federmann and William D. Lewis. 2016. Microsoft speech language translation (MSLT) corpus: The IWSLT 2016 release for English, French and German. In *International Workshop on Spoken Language Translation (IWSLT)*. Seattle, USA.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pages 44–52.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pages 128–135.

Pascale Fung and Percy Cheung. 2004. Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-comparable Corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, Geneva, Switzerland, COLING '04. https://doi.org/10.3115/1220355.1220506.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-Lingual Discriminative Learning of Sequence Models with Posterior Regularization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1996–2006.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The Impact of Sentence Alignment Errorson Phrase-Based Machine Translation Performance. In *Proceedings of AMTA-2012: The Tenth Biennial Conference of the Association for Machine Translation in the Americas*.

Nizar Habash and Bonnie Dorr. 2002. Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In *Conference of the Association for Machine Translation in the Americas*. Springer, pages 84–93.

Silja Hildebrand and Stephan Vogel. 2013. MT Quality Estimation: The CMU System for WMT'13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. pages 373–379.

Jie Jiang, Andy Way, and Julie Carson-Berndsen. 2010. Lattice score based data cleaning for phrase-based statistical machine translation. In *14th Annual Conference of the European Association for Machine Translation (EAMT)*.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality + ML: Learning Adaptive Similarity Functions for Cross-lingual Textual Entailment. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics, Montreal, Canada, SemEval '12, pages 684–688.

Alon Lavie. 2008. Stat-XFER: A general search-based syntax-driven framework for machine translation. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 362–375.

William Lewis, Christian Federmann, and Ying Xin. 2016. Applying Cross-Entropy Difference for Selecting Parallel Training Data from Publicly Available Sources for Conversational Machine Translation. In *International Workshop on Spoken Language Translation*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Austin Matthews, Waleed Ammar, Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2014. The CMU Machine Translation Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 142–149.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards Cross-lingual Textual Entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, HLT '10, pages 321–324.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised Attentions for Neural Machine Translation. In *Prceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting Parallel Sub-sentential Fragments from Non-parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, ACL-44, pages 81–88. https://doi.org/10.3115/1220175.1220186.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Montr\?al, Canada, SemEval '12, pages 399–407.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 25–33.

Matteo Negri and Yashar Mehdad. 2010. Creating a Bi-lingual Entailment Corpus Through Translations with Mechanical Turk: $100 for a 10-day Rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, Los Angeles, California, CSLDAMT '10, pages 212–216.

Tsuyoshi Okita, Sudip K. Naskar, and Andy Way. 2009. Noise reduction experiments in machine translation. In *ECML-PKDD, Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Bled*.

Jason Riesa and Daniel Marcu. 2012. Automatic Parallel Fragment Extraction from Noisy Data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 538–542.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: A Toolkit for Neural Machine Translation. In *Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceddings of the 1st Conference on Machine Translation*. volume 2 Shared Task Papers.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)* .

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, HLT '10, pages 403–411.

Lucia Specia, Varvara Logacheva, and Carolina Scarton. 2016. WMT16 Quality Estimation Shared Task Training and Development Data. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation* 24(1):39–50. https://doi.org/10.1007/s10590-010-9077-2.

Jörg Tiedemann. 2007. Improved sentence alignment for movie subtitles. In *Recent Advances in Natural Language Processing (RANLP)*. volume 7.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics* 23(3):377–404.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*. San Diego, HLT '01, pages 1–8. https://doi.org/10.3115/1072133.1072187.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *Procedings of the International Conference on Learning Representations (ICLR)*. Toulon, France.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *International Conference on Data Mining (ICDM)*. IEEE, pages 745–748.

Jiang Zhao, Man Lan, and Zheng-yu Niu. 2013. ECNUCS: Recognizing cross-lingual textual entailment using multiple text similarity and text difference measures. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.