Identifying Semantic Divergences in Parallel Text without Annotations

Yogarshi Vyas and **Xing Niu** and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
yogarshi@cs.umd.edu, xingniu@cs.umd.edu, marine@cs.umd.edu

Abstract

Recognizing that even correct translations are not always semantically equivalent, we automatically detect meaning divergences in parallel sentence pairs with a deep neural model of bilingual semantic similarity which can be trained for any parallel corpus without any manual annotation. We show that our semantic model detects divergences more accurately than models based on surface features derived from word alignments, and that these divergences matter for neural machine translation.

1 Introduction

Parallel sentence pairs are sentences that are translations of each other, and are therefore often assumed to convey the same meaning in the source and target language. Occasional mismatches between source and target have been primarily viewed as alignment noise (Goutte et al., 2012) due to imperfect sentence alignment tools in parallel corpora drawn from translated texts (Tiedemann, 2011; Xu and Yvon, 2016), or the noisy process of extracting parallel segments from non-parallel corpora (Fung and Cheung, 2004; Munteanu and Marcu, 2005).

In contrast, we view translation as a process that inherently introduces meaning mismatches, so that even correctly aligned sentence pairs are not necessarily semantically equivalent. This can happen for many reasons: translation lexical choice often involves selecting between near synonyms that introduce language-specific nuances (Hirst, 1995), typological divergences lead to structural mismatches (Dorr, 1994), differences in discourse organization can make it impossible to find one-to-one sentence alignments (Li et al., 2014). Cross-linguistic variations in other discourse phenomena such as coreference, discourse relation and modality (Lapshinova-Koltunski, 2015) compounded with translation effects that distinguish "translationese" from original text (Koppel and Ordan, 2011) might also lead to meaning mismatches between source and target.

In this paper, we aim to provide empirical evidence that semantic divergences exist in parallel corpora and matter for downstream applications. This requires an automatic method to distinguish semantically equivalent sentence pairs from semantically divergent pairs, so that parallel corpora can be used more judiciously in downstream cross-lingual NLP applications. We propose a semantic model to automatically detect whether a sentence pair is semantically divergent (Section 3). While prior work relied on surface cues to detect mis-aligments, our approach focuses on comparing the meaning of words and overlapping text spans using bilingual word embeddings (Luong et al., 2015) and a deep convolutional neural network (He and Lin, 2016). Crucially, training this model requires no manual annotation. Noisy supervision is obtained automatically borrowing techniques developed for parallel sentence extraction (Munteanu and Marcu, 2005). Our model can thus easily be trained to detect semantic divergences in any parallel corpus.

We extensively evaluate our semantically-motivated models on intrinsic and extrinsic tasks: detection of divergent examples in two parallel English-French data sets (Section 5), and data selection for English-French and Vietnamese-English machine translation (MT) (Section 6).The semantic models significantly outperform other methods on the intrinsic task, and help select data to train neural machine translation faster with no loss in translation quality. Taken together, these results provide empirical evidence that sentence-alignment does not necessarily imply semantic equivalence, and that this distinction matters in practice for a downstream NLP application.

Proceedings of NAACL-HLT 2018, pages 1503–1515
New Orleans, Louisiana, June 1 - 6, 2018.

2 Background

Translation Divergences We use the term semantic divergences to refer to bilingual sentence pairs, including translations, that do not have the same meaning. These differ from *typological divergences*, which have been defined as structural differences between sentences that convey the same meaning (Dorr, 1994; Habash and Dorr, 2002), and reflect the fact that languages do not encode the same information in the same way.

Non-Parallel Corpora Mismatches in bilingual sentence pairs have previously been studied to extract parallel segments from non-parallel corpora, and augment MT training data (Fung and Cheung, 2004; Munteanu and Marcu, 2005, 2006; Abdul-Rauf and Schwenk, 2009; Smith et al., 2010; Riesa and Marcu, 2012, *inter alia)*. Methods for parallel sentence extraction rely primarily on surface features based on translation lexicons and word alignment patterns (Munteanu and Marcu, 2005, 2006). Similar features have proved to be useful for the related task of translation quality estimation (Specia et al., 2010, 2016), which aims to detect divergences introduced by MT errors, rather than human translation. Recently, sentence embeddings have also been used to detect parallelism (España-Bonet et al., 2017; Schwenk and Douze, 2017). Although embeddings capture semantic generalizations, these models are trained with neural MT objectives, which do not distinguish semantically equivalent segments from divergent parallel segments.

Cross-Lingual Sentence Semantics Cross-lingual semantic textual similarity (Agirre et al., 2016) and cross-lingual textual entailment (Negri and Mehdad, 2010; Negri et al., 2012, 2013) seek to characterize semantic relations between sentences in two different languages beyond translation equivalence, and are therefore directly relevant to our goal. While the human judgments obtained for each task differ, they all take inputs of the same form (two segments in 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 transfer the cross-lingual task into its monolingual equivalent (Jimenez et al., 2012; Zhao et al., 2013), or on features derived from MT components such as translation dictionaries and word alignments (Turchi and Negri, 2013; Lo et al., 2016). Training requires manually annotated examples, either bilingual, or monolingual when using MT for language transfer.

Impact of mismatched sentence pairs on MT Prior MT work has focused on noise in sentence alignment rather than semantic divergence. Goutte et al. (2012) show that phrase-based systems are remarkably robust to noise in parallel segments. When introducing noise by permuting the target side of parallel pairs, as many as 30% of training examples had to be permuted to degrade BLEU significantly. While such artificial noise does not necessarily capture naturally occurring divergences, there is evidence that data cleaning to remove real noise can benefit MT in low-resource settings (Matthews et al., 2014).

Neural MT quality appears to be more sensitive to the nature of training examples than phrase-based systems. Chen et al. (2016) suggest that neural MT systems are sensitive to sentence pair permutations in domain adaptation settings. Belinkov and Bisk (2018) demonstrate the brittleness of character-level neural MT when exposed to synthetic noise (random permutations of words and characters) as well as natural human errors. Concurrently with our work, Hassan et al. (2018) claim that even small amounts of noise can have adverse effects on neural MT models, as they tend to assign high probabilities to rare events. They filter out noise and select relevant in-domain examples jointly, using similarities between sentence embeddings obtained from the encoder of a bidirectional neural MT system trained on clean in-domain data. In contrast, we detect semantic divergence with dedicated models that require only 5000 parallel examples (see Section 5).

This work builds on our initial study of semantic divergences (Carpuat et al., 2017), where we provide a framework for evaluating the impact of meaning mismatches in parallel segments on MT via data selection: we show that filtering out the most divergent segments in a training corpus improves translation quality. However, we previously detect mismatches using a cross-lingual entailment classifier, which is based on surface features only, and requires manually annotated training examples (Negri et al., 2012, 2013). In this paper, we detect divergences using a semantically-motivated model that can be trained given any existing parallel corpus without manual intervention.

3 Approach

We introduce our approach to detecting divergence in parallel sentences, with the goal of (1) detecting differences ranging from large mismatches to subtle nuances, (2) without manual annotation.

Cross-Lingual Semantic Similarity Model We address the first requirement using a neural model that compares the meaning of sentences using a range of granularities. We repurpose the Very Deep Pairwise Interaction (VDPWI) model, which has been previously been used to detect semantic textual similarity (STS) between English sentence pairs (He and Lin, 2016). It achieved competitive performance on data from the STS 2014 shared task (Agirre et al., 2014), and outperformed previous approaches on sentence classification tasks (He et al., 2015; Tai et al., 2015), with fewer parameters, faster training, and without requiring expensive external resources such as WordNet.

The VDPWI model was designed for comparing the meaning of sentences in the same language, based not only on word-to-word similarity comparisons, but also on comparisons between overlapping spans of the two sentences, as learned by a deep convolutional neural network. We adapt the model to our cross-lingual task by initializing it with bilingual embeddings. To the best of our knowledge, this is the first time this model has been used for cross-lingual tasks in such a way. We give a brief overview of the resulting model here and refer the reader to the original paper for details. Given sentences e and f , VDPWI models the semantic similarity between them using a pipeline consisting of five components:

1. **Bilingual word embeddings**: Each word in e and f is represented as a vector using pretrained, bilingual embeddings.
2. **BiLSTM for contextualizing words**: Contextualized representations for words in e and f are obtained by choosing the output vectors at each time step obtained by running a bidirectional LSTM (Schuster and Paliwal, 1997) on each sentence.
3. **Word similarity cube**: The contextualized representations are used to calculate various similarity scores between each word in *e* with each word in *f*. Each score thus forms a matrix and all such matrices are stacked to form a *similarity cube* tensor.
4. **Similarity focus layer**: The similarity cube is fed to a similarity focus layer that reweights the similarities in the cube to focus on highly similar word pairs, by decreasing the weights of pairs which are less similar. This output is called the *focus cube*.
5. **Deep convolutional network**: The focus cube is treated as an "image" and passed to a deep neural network, the likes of which have been used to detect patterns in images. The network consists of repeating convolution and pooling layers. Each repetition consists of a spatial convolutional layer, a Rectified Linear Unit (Nair and Hinton, 2010), and a max pooling layer, followed by fully connected layers, and a softmax to obtain the final output.

The entire architecture is trained end-to-end to minimize the Kullback-Leibler divergence (Kullback, 1959) between the output similarity score and gold similarity score.

Noisy Synthetic Supervision How can we obtain gold similarity scores as supervision for our task? We automatically construct examples of semantically divergent and equivalent sentences as follows. Since a large number of parallel sentence pairs are semantically equivalent, we use parallel sentences as positive examples. Synthetic negative examples are generated following the protocol introduced by Munteanu and Marcu (2005) to distinguish parallel from non-parallel segments. Specifically, candidate negative examples are generated starting from the positive examples $\{(e_i, f_i)\ \forall i\}$ and taking the Cartesian product of the two sides of the positive examples$\{(e_i, f_j) \forall i, j$ s.t. $i \neq j\}$. This candidate set is filtered to ensure that negative examples are not too easy to identify: we only retain pairs that are close to each other in length (a length ratio of at most 1:2), and have enough words (at least half) which have a translation in the other sentence according to a bilingual dictionary derived from automatic word alignments.

This process yields positive and negative examples that are a noisy source of supervision for our task, as some of the positive examples might not be fully equivalent in meaning. However, experiments will show that, in aggregate, they provide a useful signal for the VDPWI model to learn to detect semantic distinctions (Section 5).

		Equivalent with High Agreement ($n = 5$)
subs	en	the epidemic took my wife, my stepson.
	fr	l'épidémie a touché ma femme, mon beau-fils.
	gl	the epidemic touched my wife, my stepson.
		Equivalent with Low Agreement ($n = 3$)
cc	en	cancellation policy: if cancelled up to 28 days before date of arrival, no fee will be charged.
	fr	conditions d'annulation : en cas d'annulation jusqu'à 28 jours avant la date d'arrivée, l'hôtel ne prélève pas de frais sur la carte de crédit fournie.
	gl	cancellation conditions: in case of cancellation up to 28 days before arrival date, the hotel does not charge fees from the credit card given.
		Divergent with Low Agreement ($n = 3$)
cc	en	what does it mean when food is "low in ash" or "low in magnesium"?
	fr	quels sont les avantages dune nourriture "réduite en cendres" et "faible en magnésium" ?
	gl	what are the advantages of a food "low in ash" or "low in magnesium"?
		Divergent with High Agreement ($n = 5$)
subs	en	rabbit? if i told you it was a chicken, you wouldn't know the difference.
	fr	vous croirez manger du poulet.
	gl	you think eat chicken

Table 1: Randomly selected sentence pairs (English (en), French (fr) and gloss of French (gl)) annotated as divergent or equivalent, with high and low degrees of agreement between the 5 annotators. Examples are taken either from the OpenSubtitles (subs) or Common Crawl (cc) corpus.

4 Crowdsourcing Divergence Judgments

We crowdsource annotations of English-French sentence pairs to construct test beds for evaluating our models, and also to assess how frequent semantic divergences are in parallel corpora.

Data Selection We draw examples for annotation randomly from two English-French corpora, using a resource-rich and well-studied language pair, and for which bilingual annotators can easily be found. The **OpenSubtitles corpus** contains 33M sentence pairs based on translations of movie subtitles. The sentence pairs are expected to not be completely parallel given the many constraints imposed on translations that should fit on a screen and be synchronized with a movie (Tiedemann, 2007; Lison and Tiedemann, 2016), and the use of more informal registers which might require frequent non-literal translations of figurative language. The **Common Crawl corpus** contains sentence-aligned parallel documents automatically mined from the Internet. Parallel documents are discovered using e.g., URL containing language code patterns, and sentences are automatically aligned after structural cleaning of HTML. The resulting corpus of 3M sentence pairs is noisy, yet extremely useful to improve translation quality for multiple language pairs and domains (Smith et al., 2013).

Annotation Protocol Divergence annotations are obtained via Crowdflower.[1] Since this task requires good command of both French and English, we rely on a combination of strategies to obtain good quality annotations, including Crowdflower's internal worker proficiency ratings, geo-restriction, reference annotations by a bilingual speaker in our lab, and instructions that alternate between the two languages (Agirre et al., 2016).

Annotators are shown an English-French sentence pair, and asked whether they agree or disagree with the statement "the French and English text convey the same information." We do not use the term "divergent", and let the annotators' intuitions about what constitutes the same take precedence. We set up two distinct annotation tasks, one for each corpus, so that workers only see examples sampled from a given corpus in a given job. Each example is shown to five distinct annotators.

Annotation Analysis Forcing an assignment of divergent or equivalent labels by majority vote yields 43.6% divergent examples in OpenSubtitles, and 38.4% in Common Crawl. Fleiss' Kappa indicates moderate agreement between annotators (0.41 for OpenSubtitles and 0.49 for Common Crawl). This suggests that the annotation protocol can be improved, perhaps by using graded judgments as in Semantic Textual Similarity tasks (Agirre et al., 2016), or for sentence alignment

[1] http://crowdflower.com

confidence evaluation (Xu and Yvon, 2016). Current annotations are nevertheless useful, and different degrees of agreement reveal nuances in the nature of divergences (Table 1). Examples labeled as divergent with high confidence (lowest block of the table) are either unrelated or one language misses significant information that is present in the other. Examples labeled divergent with lower confidence contain more subtle differences (e.g., "what does it mean" in English vs. "what are the advantages" in French).

5 Divergence Detection Evaluation

Using the two test sets obtained above, we can evaluate the accuracy of our cross-lingual semantic divergence detector, and compare it against a diverse set of baselines in controlled settings. We test our hypothesis that semantic divergences are more than alignment mismatches by comparing the semantic divergence detector with models that capture mis-alignment (Section 5.2) or translation (Section 5.3). Then, we compare the deep convolutional architecture of the semantic divergence model, with a much simpler model that directly compares bilingual sentence embeddings (Section 5.4). Finally, we compare our model trained on synthetic examples with a supervised classifier used in prior work to predict finer-grained textual entailment categories based on manually created training examples (Section 5.5). Except for the entailment classifier which uses external resources, all models are trained on the exact same parallel corpora (OpenSubtitles or CommonCrawl for evaluating on the corresponding test bed.)

5.1 Neural Semantic Divergence Detection

Model and Training Settings We use the publicly available implementation of the VDPWI model.[2] We initialize with 200 dimensional BiVec French-English word embeddings (Luong et al., 2015), trained on the parallel corpus from which our test set is drawn. We use the default setting for all other VDPWI parameters. The model is trained for 25 epochs and the model that achieves the best Pearson correlation coefficient on the validation set is chosen. At test time, VDPWI outputs a score $\in [0, 1]$, where a higher value indicates less divergence. We tune a threshold on development data to convert the real-valued score to binary predictions.

Synthetic Data Generation The synthetic training data is constructed using a random sample of 5000 sentences from the training parallel corpus as positive examples. We generate negative examples automatically as described in Section 3, and sample a subset to maintain a 1:5 ratio of positive to negative examples.[3]

5.2 Parallel vs. Non-parallel Classifier

Are divergences observed in parallel corpora more than alignment errors? To answer this question, we reimplement the model proposed by Munteanu and Marcu (2005). It discriminates parallel pairs from non-parallel pairs in comparable corpora using a supervised linear classifier with the following features for each sentence pair (e, f):

- Length features: $|f|$, $|e|$, $\frac{|f|}{|e|}$, and $\frac{|e|}{|f|}$
- Alignment features (for each of e and f):[4]
  - Count and ratio of unaligned words
  - Top three largest fertilities
  - Longest contiguous unaligned and aligned sequence lengths
- Dictionary features:[5] fraction of words in e that have a translation in f and vice-versa.

Training uses the exact same synthetic examples as the semantic divergence model (Section 3).

5.3 Neural MT

If divergent examples are nothing more than bad translations, a neural MT system should assign lower scores to divergent segments pairs than to those that are equivalent in meaning. We test this empirically using neural MT systems trained for a single epoch, and use the system to score each of the sentence pairs in the test sets. We tune a threshold on the development set to convert scores to binary predictions. The system architecture and training settings are described in details in the later MT section (Section 6.2). Preliminary experiments showed that training for more than one epoch does not help divergence detection.

[2] https://github.com/castorini/VDPWI-NN-Torch

[3] We experimented with other ratios and found that the results only slightly degraded while using a more balanced ratio (1:1, 1:2), but severely degraded with a skewed ratio (1:9).

[4] Alignments are obtained using IBM Model 2 trained in each direction, combined with `union`, `intersection`, and `grow-diag-final-and` heuristics.

[5] The bilingual dictionary used to extract features is constructed using word alignments from a random sample of a million sentences from the training parallel corpus.

5.4 Bilingual Sentence Embeddings

Our semantic divergence model introduces a large number of parameters to combine the pairwise word comparisons into a single sentence-level prediction. This baseline tests whether a simpler model would suffice. We detect semantic divergence by computing the cosine similarity between sentence embeddings in a bilingual space. The sentence embeddings are bag-of-word representations, build by taking the mean of bilingual word embeddings for each word in the sentence. This approach has been shown to be effective, despite ignoring fundamental linguistic information such as word order and syntax (Mitchell and Lapata, 2010). We use the same 200 dimensional BiVec word embeddings (Luong et al., 2015), trained on OpenSubtitles and CommonCrawl respectively.

5.5 Textual Entailment Classifier

Our final baseline replicates our previous system (Carpuat et al., 2017) where we repurposed annotations and models designed for the task of Cross-Lingual Textual Entailment (CLTE) to detect semantic divergences. This baseline also helps us understand how the synthetic training data compares to training examples generated manually, for a related cross-lingual task. Using CLTE datasets from SemEval (Negri et al., 2012, 2013), we train a supervised linear classifier that can distinguish sentence pairs that entail each other, from pairs where entailment does not hold in at least one direction. The features of the classifier are based on word alignments and sentence lengths.[6]

5.6 Intrinsic Evaluation Results

Table 2 shows that the semantic similarity model is most successful at distinguishing equivalent from divergent examples. The break down per class shows that both equivalent and divergent examples are better detected. The improvement is larger for divergent examples with gains of about 10 points for F-score for the divergent class, when compared to the next-best scores.

Among the baseline methods, the non-entailment model is the weakest. While it uses manually constructed training examples, these examples are drawn from distant domains, and the categories annotated do not exactly match the task at hand. In contrast, the other models benefit from training on examples drawn from the same corpus as each test set.

Next, the machine translation based model and the sentence embedding model, both of which are unsupervised, are significantly weaker than the two supervised models trained on synthetic data, highlighting the benefits of the automatically constructed divergence examples. The strength of the semantic similarity model compared to the sentence embeddings model highlights the benefits of the fine-grained representation of bilingual sentence pairs as a similarity cube, as opposed to the bag-of-words sentence embedding representation.

Finally, despite training on the same noisy synthetic data as the parallel v/s non-parallel system, the semantic similarity model is better able to detect meaning divergences. This highlights the benefits of (1) meaning comparison between words in a shared embedding space, over the discrete translation dictionary used by the baseline, and of (2) the deep convolutional neural network which enables the semantic comparison of overlapping spans in sentence pairs, as opposed to more local word alignment features.

5.7 Analysis

We manually examine the 13-15% of examples in each test set that are correctly detected as divergent by semantic similarity and mis-classified by the non-parallel detector.

On OpenSubtitles, most of these examples are true divergences rather than noisy alignments (i.e. sentences that are not translation of each other.) The non-parallel detector weighs length features highly, and is fooled by sentence pairs of similar length that share little content and therefore have very sparse word alignments. The remaining sentence pairs are plausible translations in some context that still contain inherent divergences, such as details missing or added in one language. The non-parallel detector views these pairs as non-divergent since most words can be aligned. The semantic similarity model can identify subtle meaning differences, and correctly classify them as divergent. As a result, the non-parallel detector has a higher false positive rate (22%) than the semantic similarity classifier (14%), while having similar false negative rates (11% v/s 12%).

On the CommonCrawl test set, the examples with disagreement are more diverse, ranging from

[6] As in the prior work, alignments are trained on 2M sentence pairs from Europarl (Koehn, 2005) using the Berkeley aligner (Liang et al., 2006). The classifier is the linear SVM from Scikit-Learn.

Divergence Detection Approach	OpenSubtitles +P	+R	+F	-P	-R	-F	Overall F	Common Crawl +P	+R	+F	-P	-R	-F	Overall F
Sentence Embeddings	65	60	62	56	61	58	60	78	58	66	52	**74**	61	64
MT Scores (1 epoch)	67	53	59	54	68	60	60	54	65	59	17	11	14	42
Non-entailment	58	78	66	53	30	38	54	73	49	58	48	72	57	58
Non-parallel	70	83	76	61	42	50	66	70	83	76	61	42	49	67
Semantic Dissimilarity	**76**	**80**	**78**	**75**	**70**	**72**	**77**	**82**	**88**	**85**	**78**	69	**73**	**80**

Table 2: Intrinsic evaluation on crowdsourced semantic equivalence vs. divergence testsets. We report overall F-score, as well as precision (P), recall (R) and F-score (F) for the equivalent (+) and divergent (-) classes separately. Semantic similarity yields better results across the board, with larger improvements on the divergent class.

noisy segments that should not be aligned to sentences with subtle divergences.

6 Machine Translation Evaluation

Having established the effectiveness of the semantic divergence detector, we now measure the impact of divergences on a downstream task, machine translation. As in our prior work (Carpuat et al., 2017), we take a data selection approach, selecting the least divergent examples in a parallel corpus based on a range of divergence detectors, and comparing the translation quality of the resulting neural MT systems.

6.1 Translation Tasks

English-French We evaluate on 4867 sentences from the Microsoft Spoken Language Translation dataset (Federmann and Lewis, 2016) as well as on 1300 sentences from TED talks (Cettolo et al., 2012), as in past work (Carpuat et al., 2017). Training examples are drawn from OpenSubtitles, which contains ~28M examples after deduplication. We discard 50% examples for data selection.

Vietnamese-English Since the SEMANTIC SIMILARITY model was designed to be easily portable to new language pairs, we also test its impact on the IWSLT Vietnamese-English TED task, which comes with ~120,000 and 1268 in-domain sentences for training and testing respectively (Farajian et al., 2016). This is a more challenging translation task as Vietnamese and English are more distant languages, there is little training data, and the sentence pairs are expected to be cleaner and more parallel than those from OpenSubtitles. In these lower resource settings, we discard 10% of examples for data selection.

6.2 Neural MT System

We use the attentional encoder-decoder model (Bahdanau et al., 2015) implemented in the Sock-Eye toolkit (Hieber et al., 2017). Encoders and decoders are single-layer GRUs (Cho et al., 2014) with 1000 hidden units. Source and target word embeddings have size 512. Using byte-pair encoding (Sennrich et al., 2016), the vocabulary size is 50000. Maximum sequence length is set to 50.

We optimize the standard cross-entropy loss using Adam (Kingma and Ba, 2014), until validation perplexity does not decrease for 8 checkpoints. The learning rate is set to 0.0003 and is halved when the validation perplexity does not decrease for 3 checkpoints. The batch size is set to 80. At decoding time, we construct a new model by averaging the parameters for the 8 checkpoints with best validation perplexity, and decode with a beam of 5. All experiments are run thrice with distinct random seeds.

Note that the baseline in this work is much stronger than in our prior work (>5 BLEU). This is due to multiple factors that have been recommended as best practices for neural MT and have been incorporated in the present baseline – deduplication of the training data, ensemble decoding using multiple random runs, use of Adam as the optimizer instead of AdaDelta (Bahar et al., 2017; Denkowski and Neubig, 2017), and checkpoint averaging (Bahar et al., 2017) – as well as a more recent neural modeling toolkit.

6.3 English-French Results

We train English-French neural MT systems by selecting the least divergent half of the training corpus with the following criteria:

- SEMANTIC SIMILARITY (Section 3)
- PARALLEL: the non-parallel sentence detector (Section 5.2)
- ENTAILMENT: the entailment classifier (Section 5.5), as in Carpuat et al. (2017)
- RANDOM: Randomly downsampling the training corpus

Learning curves (Figure 1) show that data selected using SEMANTIC SIMILARITY yields better

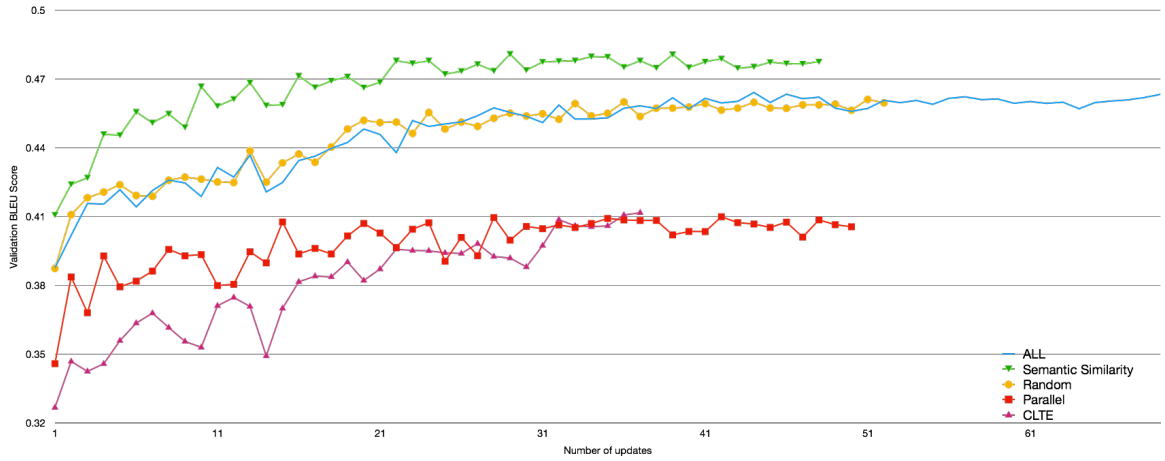

Figure 1: Learning curves on the validation set for English-French models (mean of 3 runs/model). The SEMANTIC SIMILARITY model outperforms other models throughout training, including the one trained on all data.

Model	MSLT BLEU		TED BLEU	
	Avg.	Ensemble	Avg.	Ensemble
RANDOM	43.49	45.64	36.05	38.20
PARALLEL	40.65	42.12	35.99	37.86
ENTAILMENT	39.64	41.86	33.30	35.40
SEMANTIC SIM.	**45.53**	**47.23***	**36.98**	**38.87**
ALL	44.64	46.26	36.98	38.59

Table 3: English-French decoding results. BLEU scores are either averaged across 3 runs ("Avg.") or obtained via ensemble decoding ("Ensemble"). SEMANTIC SIMILARITY reaches BLEU scores on par with ALL with only half of the training data. SEMANTIC SIMILARITY scores marked with * are significanly better ($p < 0.05$) than the corresponding ALL scores.

validation BLEU throughout training compared to all other models. SEMANTIC SIMILARITY selects more useful examples for MT than PARALLEL, even though both selection models are trained on the same synthetic examples. This highlights the benefits of semantic modeling over surface mis-alignment features. Furthermore, SEMANTIC SIMILARITY achieves the final validation BLEU of the model that uses ALL data with only 30% of the updates. This suggests that semantically divergent examples are pervasive in the training corpus, confirming the findings from manual annotation (Section 4), and that the presence of such examples slows down neural MT training.

Decoding results on the blind test sets (Table 3) show that SEMANTIC SIMILARITY outperforms all other data selection criteria (with differences being statistically significant ($p < 0.05$) (Koehn, 2004)), and performs as well or better than the ALL model which has access to twice as many training examples.

Model	Avg. Test Set BLEU
RANDOM (90%)	22.71
SEMANTIC SIM. (90%)	**23.38**
ALL	23.30

Table 4: Vietnamese-English decoding results: dropping 10% of the data based on SEMANTIC SIMILARITY does not hurt BLEU, and performs significantly ($p < 0.05$) better than RANDOM selection.

The SEMANTIC SIMILARITY model also achieves significantly better translation quality than the ENTAILMENT model used in our prior work. Surprisingly, the ENTAILMENT model performs worse than the ALL baseline, unlike in our prior work. We attribute this different behavior to several factors: the strength of the new baseline (Section 6.2), the use of Adam instead of AdaDelta, which results in stronger BLEU scores at the beginning of the learning curves for all models, and finally the deduplication of the training data. In our prior systems, the training corpus was not deduplicated. Data selection had a side-effect of reducing the ratio of duplicated examples. When the ENTAILMENT model was used, longer sentence pairs with more balanced length were selected, yielding longer translations with a better BLEU brevity penalty than the baseline. With the new systems, these advantages vanish. We further analyze output lengths in Section 6.5.

6.4 Vietnamese-English Results

Trends from English-French carry over to Vietnamese English, as the SEMANTIC SIMILARITY model compares favorably to ALL while reducing the number of training updates by 10%. SEMAN-

TIC SIMILARITY also yields better BLEU than RANDOM with the differences being statistically significant. While differences in score here are smaller, these result are encouraging since they demonstrate that our semantic divergence models port to more distant low-resource language pairs.

6.5 Analysis

We break down the results seen in Figure 1 and Table 3, with a focus on the behavior of the ENTAILMENT and ALL models. We start by analyzing the BLEU brevity penalty trends observed on the validation set during training (Figure 2).

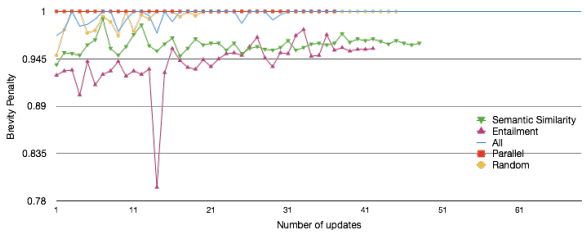

Figure 2: Brevity penalties on the validation set for English-French models (single run).

We observe that both the ENTAILMENT and SEMANTIC SIMILARITY based models have similar brevity penalties despite having performances that are at opposite ends of the spectrum in terms of BLEU. This implies that translations generated by the SEMANTIC SIMILARITY model have better n-gram overlap with the reference, but are much shorter. Manual examination of the translations suggests that the ENTAILMENT model often fails by under-translating sentences, either dropping segments from the beginning or the end of source sentences (Table 5).

The PARALLEL model consistently produces translations that are longer than the reference. [7] This is partially due to the model's propensity to generate a sequence of garbage tokens in the beginning of a sentence, especially while translating shorter sentences. In our test set, almost 12% of the translated sentences were found to begin with the garbage text shown in Table 5. Only a small fraction ($< 0.02\%$) of the French sentences in our training data begin with these tokens, but the tendency of PARALLEL to promote divergent examples above non-divergent ones, seems to exaggerate the generation of this sequence.

[7] The brevity penalty does not penalize translations that are longer than the reference.

ENTAILMENT is inadequate due to under-translation	
Source	he's a very impressive man **and still goes out to do digs.**
Reference	c'est un homme très impressionnant **et il fait encore des fouilles.**
ENTAILMENT	c'est un homme très impressionnant.
Source	when the Heat **first** won.
Reference	lorsque les Heat ont gagné **pour la première fois**.
ENTAILMENT	quand le Heat a gagné.
PARALLEL produces garbage tokens	
Source	alright.
Reference	d'accord.
ENTAILMENT	{ \ **pos (192,210)**} d'accord.

Table 5: Selected translation examples from the ensemble systems of the various models.

7 Conclusion

We conducted an extensive empirical study of semantic divergences in parallel corpora. Our crowdsourced annotations confirms that correctly aligned sentences are not necessarily meaning equivalent. We introduced an approach based on neural semantic similarity that detects such divergences much more accurately than shallower translation or alignment based models. Importantly, our model does not require manual annotation, and can be trained for any language pair and domain with a parallel corpus. Finally, we show that filtering out divergent examples helps speed up the convergence of neural machine translation training without loss in translation quality, for two language pairs and data conditions. New datasets and models introduced in this work are available at `http://github.com/yogarshi/SemDiverge`.

These findings open several avenues for future work: How can we improve divergence detection further? Can we characterize the nature of the divergences beyond binary predictions? How do divergent examples impact other applications, including cross-lingual NLP applications and semantic models induced from parallel corpora, as well as tools for human translators and second language learners?

Acknowledgments

We thank the CLIP lab at the University of Maryland and the anonymous reviewers from NAACL 2018 and WMT 2017 for their constructive feedback. This work was supported in part by research awards from Amazon, Google, and the Clare Boothe Luce Foundation.

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, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics Dublin, Ireland, pages 81–91.

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.

Parnia Bahar, Tamer Alkhouli, Jan-Thorsten Peter, Christopher Jan-Steffen Brix, and Hermann Ney. 2017. Empirical investigation of optimization algorithms in neural machine translation. *The Prague Bulletin of Mathematical Linguistics* 108(1):13–25.

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)*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*. Vancouver, Canada.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, pages 69–79. http://aclweb.org/anthology/W17-3209.

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.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of EMNLP 2014*. ArXiv: 1406.1078. http://arxiv.org/abs/1406.1078.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, pages 18–27. http://aclweb.org/anthology/W17-3203.

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

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *arXiv:1704.05415 [cs]* .

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.

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.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The Impact of Sentence Alignment Errors on 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.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt,

William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv:1803.05567 [cs]* https://arxiv.org/pdf/1803.05567.pdf.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1576–1586.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of NAACL-HLT*. pages 937–948.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *ArXiv e-prints* https://arxiv.org/abs/1712.05690.

Graeme Hirst. 1995. Near-synonymy and the structure of lexical knowledge. In *AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*. pages 51–56.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. ArXiv: 1412.6980. http://arxiv.org/abs/1412.6980.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*. http://www.aclweb.org/anthology/W04-3250.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*. Phuket, Thailand.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 1318–1326. http://dl.acm.org/citation.cfm?id=2002472.2002636.

Solomon Kullback. 1959. *Information theory and statistics*. Courier Corporation.

Ekaterina Lapshinova-Koltunski. 2015. Exploration of inter-and intralingual variation of discourse phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. pages 158–167.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 283–288.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-NAACL '06, pages 104–111. https://doi.org/10.3115/1220835.1220849.

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*.

Chi-kiu Lo, Cyril Goutte, and Michel Simard. 2016. CNRC at SemEval-2016 Task 1: Experiments in Crosslingual Semantic Textual Similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 668–673.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. pages 151–159.

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.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science* 34(8):1388–1429.

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.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omnipress, USA, ICML'10, pages 807–814.

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 Bilingual 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.

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.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pages 157–167. http://aclweb.org/anthology/W17-2619.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of 51st Meeting of the Association for Computational Linguistics*. pages 1374–1383.

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.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of ACL 2015*. Beijing, China.

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

Jörg Tiedemann. 2011. *Bitext Alignment*, volume 4.

Marco Turchi and Matteo Negri. 2013. ALTN: Word Alignment Features for Cross-lingual Textual Entailment. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Yong Xu and François Yvon. 2016. Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia, page 10.

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)*.