Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank

Eleftheria Briakou and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
ebriakou@cs.umd.edu, marine@cs.umd.edu

Abstract

Detecting fine-grained differences in content conveyed in different languages matters for cross-lingual NLP and multilingual corpora analysis, but it is a challenging machine learning problem since annotation is expensive and hard to scale. This work improves the prediction and annotation of fine-grained semantic divergences. We introduce a training strategy for multilingual BERT models by learning to rank synthetic divergent examples of varying granularity. We evaluate our models on the **R**ationalized **E**nglish-**Fr**ench **S**emantic **D**ivergences, a new dataset released with this work, consisting of English-French sentence-pairs annotated with semantic divergence classes and token-level rationales. Learning to rank helps detect fine-grained sentence-level divergences more accurately than a strong sentence-level similarity model, while token-level predictions have the potential of further distinguishing between coarse and fine-grained divergences.

1 Introduction

Comparing and contrasting the meaning of text conveyed in different languages is a fundamental NLP task. It can be used to curate clean parallel corpora for downstream tasks such as machine translation (Koehn et al., 2018), cross-lingual transfer learning, or semantic modeling (Ganitkevitch et al., 2013; Conneau and Lample, 2019), and it is also useful to directly analyze multilingual corpora. For instance, detecting the commonalities and divergences between sentences drawn from English and French Wikipedia articles about the same topic would help analyze language bias (Bao et al., 2012; Massa and Scrinzi, 2012), or mitigate differences in coverage and usage across languages (Yeung et al., 2011; Wulczyn et al., 2016; Lemmerich et al., 2019). This requires not only detecting coarse content mismatches, but also fine-grained differences in sentences that overlap in content. Consider the following English and French sentences, sampled from the WikiMatrix parallel corpus. While they share important content, highlighted words convey meaning missing from the other language:

> EN *Alexander Muir's "The Maple Leaf Forever"* ***served for many years*** *as an* ***unofficial*** *Canadian* ***national anthem***.
> FR *Alexander Muir compose The Maple Leaf Forever (en) qui est un* ***chant patriotique*** *pro canadien* ***anglais***.
> GLOSS *Alexander Muir composes The Maple Leaf Forever which is an English Canadian patriotic song.*

We show that explicitly considering diverse types of semantic divergences in bilingual text benefits both the annotation and prediction of cross-lingual semantic divergences. We create and release the **R**ationalized **E**nglish-**Fr**ench **S**emantic **D**ivergences corpus (REFRESD), based on a novel divergence annotation protocol that exploits rationales to improve annotator agreement. We introduce Divergent mBERT, a BERT-based model that detects fine-grained semantic divergences without supervision by learning to rank synthetic divergences of varying granularity. Experiments on REFRESD show that our model distinguishes semantically equivalent from divergent examples much better than a strong sentence similarity baseline and that unsupervised token-level divergence tagging offers promise to refine distinctions among divergent instances. We make our code and data publicly available.[1]

[1]Implementations of Divergent mBERT can be found at: https://github.com/Elbria/xling-SemDiv; the REFRESD dataset is hosted at: https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1563–1580,
November 16–20, 2020.

2 Background

Following Vyas et al. (2018), we use the term **cross-lingual semantic divergences** to refer to differences in meaning between sentences written in two languages. Semantic divergences differ from typological divergences that reflect different ways of encoding the same information across languages (Dorr, 1994). In sentence pairs drawn from comparable documents—written independently in each language but sharing a topic—sentences that contain translated fragments are rarely exactly equivalent (Fung and Cheung, 2004; Munteanu and Marcu, 2005), and sentence alignment errors yield coarse mismatches in meaning (Goutte et al., 2012). In translated sentence pairs, differences in discourse structure across languages (Li et al., 2014) can lead to sentence-level divergences or discrepancies in translation of pronouns (Lapshinova-Koltunski and Hardmeier, 2017; Šoštarić et al., 2018); translation lexical choice requires selecting between near synonyms that introduce language-specific nuances (Hirst, 1995); typological divergences lead to structural mismatches (Dorr, 1994), and non-literal translation processes can lead to semantic drifts (Zhai et al., 2018).

Despite this broad spectrum of phenomena, recent work has effectively focused on coarse-grained divergences: Vyas et al. (2018) work on subtitles and Common Crawl corpora where sentence alignment errors abound, and Pham et al. (2018) focus on fixing divergences where content is appended to one side of a translation pair. By contrast, Zhai et al. (2018, 2019) introduce token-level annotations that capture the meaning changes introduced by human translators during the translation process (Molina and Hurtado Albir, 2002). However, this expensive annotation process does not scale easily.

When processing bilingual corpora, any meaning mismatches between the two languages are primarily viewed as noise for the downstream task. In shared tasks for filtering web-crawled parallel corpora (Koehn et al., 2018, 2019), the best performing systems rely on translation models, or cross-lingual sentence embeddings to place bilingual sentences on a clean to noisy scale (Junczys-Dowmunt, 2018; Sánchez-Cartagena et al., 2018; Lu et al., 2018; Chaudhary et al., 2019). When mining parallel segments in Wikipedia for the WikiMatrix corpus (Schwenk et al., 2019), examples are ranked using the LASER score (Artetxe and Schwenk, 2019), which computes cross-lingual similarity in a language-agnostic sentence embedding space. While this approach yields a very useful corpus of 135M parallel sentences in 1,620 language pairs, we show that LASER fails to detect many semantic divergences in WikiMatrix.

3 Unsupervised Divergence Detection

We introduce a model based on multilingual BERT (MBERT) to distinguish divergent from equivalent sentence-pairs (Section 3.1). In the absence of annotated training data, we derive synthetic divergent samples from parallel corpora (Section 3.2) and train via learning to rank to exploit the diversity and varying granularity of the resulting samples (Section 3.3). We also show how our model can be extended to label tokens within sentences (Section 3.4).

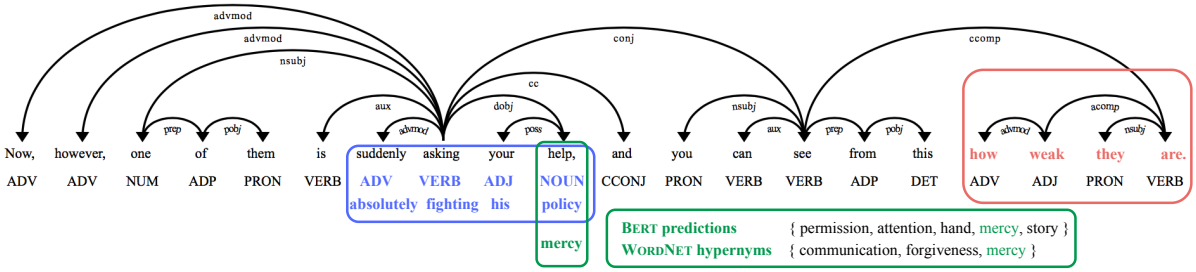
3.1 Divergent MBERT Model

Following prior work (Vyas et al., 2018), we frame divergence detection as binary classification (equivalence vs. divergence) given two inputs: an English sentence $\mathbf{x_e}$ and a French sentence $\mathbf{x_f}$. Given the success of multilingual masked language models like MBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020) on cross-lingual understanding tasks, we build our classifier on top of multilingual BERT in a standard fashion: we create a sequence $\mathbf{x}$ by concatenating $\mathbf{x_e}$ and $\mathbf{x_f}$ with helper delimiter tokens: $\mathbf{x} = ([\text{CLS}], \mathbf{x_e}, [\text{SEP}], \mathbf{x_f}, [\text{SEP}])$. The [CLS] token encoding serves as the representation for the sentence-pair $\mathbf{x}$, passed through a feed-forward layer network F to get the score $F(\mathbf{x})$. Finally, we convert the score $F(\mathbf{x})$ into the probability of $\mathbf{x}$ belonging to the equivalent class.

3.2 Generating Synthetic Divergences

We devise three ways of creating training instances that mimic divergences of varying granularity by perturbing seed equivalent samples from parallel corpora (Table 1):

Subtree Deletion We mimic semantic divergences due to content included only in one language by deleting a randomly selected subtree in the dependency parse of the English sentence, or French words aligned to English words in that subtree. We use subtrees that are not leaves, and that cover less than half of the sentence length. Durán et al. (2014); Cardon and Grabar (2020) success-



Seed Equivalent Sample

Now, however, one of them is suddenly asking your help, and you can see from this how weak they are.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles.

Subtree Deletion

Now, however, one of them is suddenly asking your help, and you can see from this.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles .

Phrase Replacement

Now, however, one of them is absolutely fighting his policy , and you can see from this how weak they are.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles.

Lexical Substitution

Now, however, one of them is suddenly asking your mercy , and you can see from this.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles.

Table 1: Starting from a seed equivalent parallel sentence-pair, we create three types of divergent samples of varying granularity by introducing the highlighted edits.

fully use this approach to compare sentences in the same language.

Phrase Replacement Following Pham et al. (2018), we introduce divergences that mimic phrasal edits or mistranslations by substituting random source or target sequences by another sequence of words with matching POS tags (to keep generated sentences as grammatical as possible).

Lexical Substitution We mimic *particularization* and *generalization* translation operations (Zhai et al., 2019) by substituting English words with hypernyms or hyponyms from WordNet. The replacement word is the highest scoring WordNet candidate in context, according to a BERT language model (Zhou et al., 2019; Qiang et al., 2019).

We call all these divergent examples **contrastive** because each divergent example contrasts with a specific equivalent sample from the seed set. The three sets of transformation rules above create divergences of varying granularity and create an **implicit ranking over divergent examples** based on the range of edit operations, starting from a single token with lexical substitution, to local short phrases for phrase replacement, and up to half the words in a sentence when deleting subtrees.

3.3 Learning to Rank Contrastive Samples

We train the Divergent mBERT model by learning to rank synthetic divergences. Instead of treating equivalent and divergent samples independently, we exploit their contrastive nature by explicitly pairing divergent samples with their seed equivalent sample when computing the loss. Intuitively, *lexical substitution* samples should rank higher than *phrase replacement* and *subtree deletion* and lower than seed *equivalents*: we exploit this intuition by enforcing a margin between the scores of increasingly divergent samples.

Formally, let $\mathbf{x}$ denote an English-French sentence-pair and $\mathbf{y}$ a contrastive pair, with $\mathbf{x} > \mathbf{y}$ indicating that the divergence in $\mathbf{x}$ is finer-grained than in $\mathbf{y}$. For instance, we assume that $\mathbf{x} > \mathbf{y}$ if $\mathbf{x}$ is generated by lexical substitution and $\mathbf{y}$ by subtree deletion.

At training time, given a set of contrastive pairs $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$, the model is trained to rank the score of the first instance higher than the latter by minimizing the following margin-based loss

$$\mathcal{L}_{\text{sent}} = \frac{1}{|\mathcal{D}|}\Big(\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \max\{0, \xi - F(\mathbf{x}) + F(\mathbf{y})\}\Big) \quad (1)$$

where ξ is a hyperparameter margin that controls the score difference between the sentence-pairs $\mathbf{x}$

and $\mathbf{y}$. This ranking loss has proved useful in supervised English semantic analysis tasks (Li et al., 2019), and we show that it also helps with our cross-lingual synthetic data.

3.4 Divergent mBERT for Token Tagging

We introduce an extension of Divergent mBERT which, given a bilingual sentence pair, produces a) a sentence-level prediction of equivalence vs. divergence and b) a sequence of EQ/DIV labels for each input token. EQ and DIV refer to token-level tags of equivalence and divergence, respectively.

Motivated by annotation rationales, we adopt a multi-task framework to train our model on a set of triplets $\mathcal{D}' = \{(\mathbf{x}, \mathbf{y}, \mathbf{z})\}$, still using only synthetic supervision (Figure 1). As in Section 3.3, we assume $\mathbf{x} > \mathbf{y}$, while $\mathbf{z}$ is the sequence of labels for the second encoded sentence pair $\mathbf{y}$, such that, at time t, $z_t \in$ {EQ,DIV} is the label of y_t. Since Divergent mBERT operates on sequences of subwords, we assign an EQ or DIV label to a word token if at leat one of its subword units is assigned that label.

For the token prediction task, the final hidden state h_t of each y_t token is passed through a feed-forward layer and a softmax layer to produce the probability P_{y_t} of the y_t token belonging to the EQ class. For the sentence task, the model learns to rank $\mathbf{x} > \mathbf{y}$, as in Section 3.3. We then minimize the sum of the sentence-level margin-loss and the average token-level cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$) across all tokens of $\mathbf{y}$, as defined in Equation 2.

$$\mathcal{L} = \frac{1}{|\mathcal{D}'|}\Big(\sum_{(\mathbf{x},\mathbf{y},\mathbf{z})\in\mathcal{D}'} \big(\max\{0, \xi - F(\mathbf{x}) + F(\mathbf{y})\} + \frac{1}{|\mathbf{y}|}\sum_{t=1}^{|\mathbf{y}|} \mathcal{L}_{\mathrm{CE}}(\mathrm{P}_{y_t}, z_t)\big)\Big) \quad (2)$$

Similar multi-task models have been used for Machine Translation Quality Estimation (Kim et al., 2019a,b), albeit with human-annotated training samples and a standard cross-entropy loss for both word-level and sentence-level sub-tasks.

4 Rationalized English-French Semantic Divergences

We introduce the **R**ationalized **E**nglish-**Fr**ench **S**emantic **D**ivergences (REFRESD) dataset, which consists of 1,039 English-French sentence-pairs annotated with sentence-level divergence judgments and token-level rationales. Figure 2 shows an example drawn from our corpus.

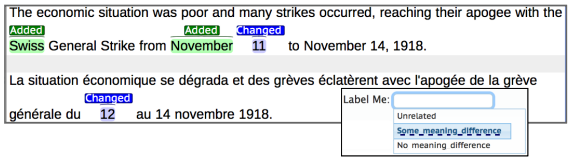


Figure 2: Screenshot of an example annotated instance.

Our annotation protocol is designed to encourage annotators' sensitivity to semantic divergences other than misalignments, without requiring expert knowledge beyond competence in the languages of interest. We use two strategies for this purpose: (1) we explicitly introduce distinct divergence categories for unrelated sentences and sentences that overlap in meaning; and (2) we ask for annotation rationales (Zaidan et al., 2007) by requiring annotators to highlight tokens indicative of meaning differences in each sentence-pair. Thus, our approach strikes a balance between coarsely annotating sentences with binary distinctions that are fully based on annotators' intuitions (Vyas et al., 2018), and exhaustively annotating all spans of a sentence-pair with fine-grained labels of translation processes (Zhai et al., 2018). We describe the annotation process and analysis of the collected instances based on data statements protocols described in Bender and Friedman (2018); Gebru et al. (2018). We include more information in A.4.

Task Description An annotation instance consists of an English-French sentence-pair. Bilingual participants are asked to read them both and highlight tokens in each sentence that convey meaning not found in the other language. For each highlighted span, they pick whether this span conveys added information ("Added"), information that is present in the other language but not an exact match ("Changed"), or some other type ("Other"). Those fine-grained classes are added to improve consistency across annotators and encourage them to read and compare the text closely. Finally, participants are asked to make a sentence-level judgment by selecting one of the following classes: "No meaning difference", "Some meaning difference", "Unrelated". Participants are not given specific instructions on how to use span annotations to make sentence-level decisions. Furthermore, participants have the option of using a text box to provide any comments or feedback on the example and their decisions. A summary of the different span and sentence labels along with the instructions given to participants can be found in A.3.

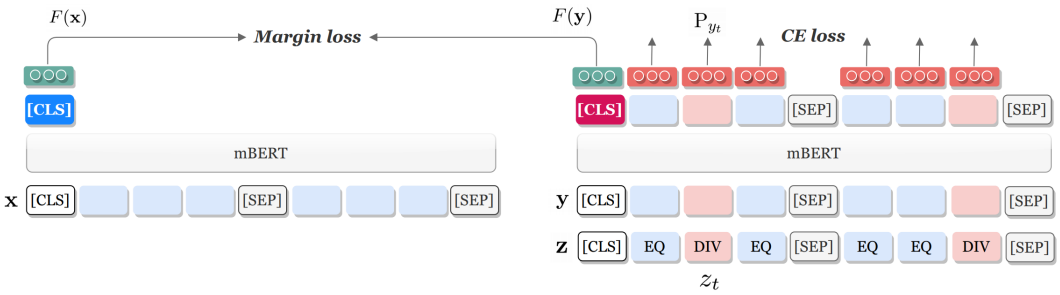


Figure 1: Divergent mBERT training strategy: given a triplet (**x**,**y**,**z**), the model minimizes the sum of a margin-based loss via ranking a contrastive pair $\mathbf{x} > \mathbf{y}$ and a token-level cross-entropy loss on sequence labels **z**.

Curation rationale Examples are drawn from the English-French section of the WikiMatrix corpus (Schwenk et al., 2019). We choose this resource because (1) it is likely to contain diverse, interesting divergence types, since it consists of mined parallel sentences of diverse topics which are not necessarily generated by (human) translations, and (2) Wikipedia and WikiMatrix are widely used resources to train semantic representations and perform cross-lingual transfer in NLP. We exclude obviously noisy samples by filtering out sentence-pairs that a) are too short or too long, b) consist mostly of numbers, c) have a small token-level edit difference. The filtered version of the corpus consists of 2,437,108 sentence-pairs.

Quality Control We implement quality control strategies at every step. We build a dedicated task interface using the BRAT annotation toolkit (Stenetorp et al., 2012) (Figure 2). We recruit participants from an educational institution and ensure they are proficient in both languages of interest. Specifically, participants are either bilingual speakers or graduate students pursuing a Translation Studies degree. We run a pilot study were participants annotate a sample containing both duplicated and reference sentence-pairs previously annotated by one of the authors. All annotators are found to be internally consistent on duplicated instances and agree with the reference annotations more than 60% of the time. We solicit feedback from participants to finalize the instructions.

Inter-annotator Agreement (IAA) We compute IAA for sentence-level annotations, as well as for the token and span-level rationales (Table 2). We report 0.60 Krippendorf's α coefficient for sentence classes, which indicates a "moderate" agreement between annotators (Landis and Koch, 1977). This constitutes a significant improvement over the 0.41 and 0.49 reported agreement coefficients on crowdsourced annotations of equivalence vs. divergence English-French parallel sentences drawn from OpenSubtitles and CommonCrawl corpora by prior work (Vyas et al., 2018).

Disagreements mainly occur between the "No meaning difference" and "Some meaning difference" classes, which we expect as different annotators might draw the line between which differences matter differently. We only observed 3 examples where all 3 annotators disagreed (tridisagreements), which indicates that the "Unrelated" and "No meaning difference" categories are more clear-cut. The rare instances with tridisagreements and bidisagreements—where the disagreement spans the two extreme classes—were excluded from the final dataset. Examples of REFRESD corresponding to different levels of IAA are included in A.5.

Granularity	Method	IAA
Sentence	Krippendorf's α	0.60
Span	macro F1	45.56 ± 7.60
Token	macro F1	33.94 ± 8.24

Table 2: Inter-annotator agreement measured at different levels of granularities for the REFRESD dataset.

Quantifying agreement between rationales requires different metrics. At the span-level, we compute macro F1 score for each sentence-pair following DeYoung et al. (2020), where we treat one set of annotations as the reference standard and the other set as predictions. We count a prediction as a match if its token-level Intersection Over Union (IOU) with any of the reference spans overlaps by more than some threshold (here, 0.5). We report average span-level and token-level macro F1 scores, computed across all different pairs of annotators. Average statistics indicate that our annotation protocol enabled the collection of a high-quality dataset.

Dataset Statistics Sentence-level annotations were aggregated by majority vote, yielding 252, 418, and 369 instances for the "Unrelated", "Some meaning difference", and "No meaning difference" classes, respectively. In other words, 64% of samples are divergent and 40% of samples contain fine-grained meaning divergences, confirming that divergences vary in granularity and are too frequent to be ignored even in a corpus viewed as parallel.

5 Experimental Setup

Data We normalize English and French text in WikiMatrix consistently using the Moses toolkit (Koehn et al., 2007), and tokenize into subword units using the "BertTokenizer". Specifically, our pre-processing pipeline consists of a) replacement of Unicode punctuation, b) normalization of punctuation, c) removing of non-printing characters, and d) tokenization.[2] We align English to French bitext using the Berkeley word aligner.[3] We filter out obviously noisy parallel sentences, as described in Section 4, Curation Rationale. The top 5,500 samples ranked by LASER similarity score are treated as (noisy) equivalent samples and seed the generation of synthetic divergent examples.[4] We split the seed set into 5,000 training instances and 500 development instances consistently across experiments. Results on development sets for each experiment are included in A.7.

Models Our models are based on the HuggingFace transformer library (Wolf et al., 2019).[5] We fine-tune the "BERT-Base Multilingual Cased" model (Devlin et al., 2019),[6] and perform a grid search on the margin hyperparameter, using the synthetic development set. Further details on model and training settings can be found in A.1.

Evaluation We evaluate all models on our new REFRESD dataset using Precision, Recall, F1 for each class, and Weighted overall F1 score as computed by *scikit-learn* (Pedregosa et al., 2011).[7]

[2] https://github.com/facebookresearch/XLM/blob/master/tools/tokenize.sh

[3] https://code.google.com/archive/p/berkeleyaligner

[4] https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix

[5] https://github.com/huggingface/transformers

[6] https://github.com/google-research/bert

[7] https://scikit-learn.org

6 Binary Divergence Detection

We evaluate Divergent mBERT's ability to detect divergent sentence pairs in REFRESD.

6.1 Experimental Conditions

LASER baseline This baseline distinguishes equivalent from divergent samples via a threshold on the LASER score. We use the same threshold as Schwenk et al. (2019), who show that training Neural Machine Translation systems on WikiMatrix samples with LASER scores higher than 1.04 improves BLEU. Preliminary experiments suggest that tuning the LASER threshold does not improve classification and that more complex models such as the VDPWI model used by Vyas et al. (2018) underperform Divergent mBERT, as discussed in A.2.

Divergent mBERT We compare Divergent mBERT trained by learning to rank contrastive samples (Section 3.3) with ablation variants.

To test the impact of contrastive training samples, we fine-tune Divergent mBERT using 1. the Cross-Entropy (**CE**) loss on randomly selected synthetic divergences; 2. the **CE** loss on paired equivalent and divergent samples, treated as independent; 3. the proposed training strategy with a **Margin** loss to explicitly compare contrastive pairs.

Given the fixed set of seed equivalent samples (Section 5, Data), we vary the combinations of divergent samples: 1. **Single divergence type** we pair each seed equivalent with its corresponding divergent of that type, yielding a single contrastive pair; 2. **Balanced sampling** we randomly pair each seed equivalent with one of its corresponding divergent types, yielding a single contrastive pair; 3. **Concatenation** we pair each seed equivalent with one of each synthetic divergence type, yielding four contrastive pairs; 4. **Divergence ranking** we learn to rank pairs of close divergence types: equivalent vs. lexical substitution, lexical substitution vs. phrase replacement, or subtree deletion yielding four contrastive pairs.[8]

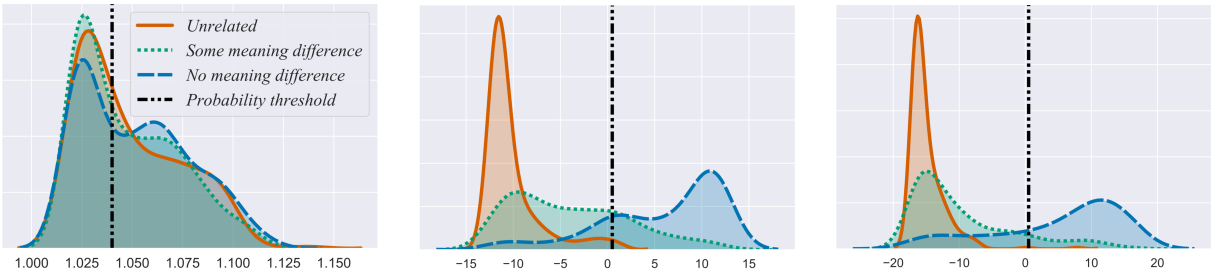
6.2 Results

All Divergent mBERT models outperform the LASER baseline by a large margin (Table 3). The proposed training strategy performs best, improving over LASER by 31 F1 points. Ablation experiments and analysis further show the benefits of diverse contrastive samples and learning to rank.

[8] We mimic both generalization **and** particularization.

Synthetic	Loss	Contrastive	*Equivalents* P+	*Equivalents* R+	*Equivalents* F1+	*Divergents* P-	*Divergents* R-	*Divergents* F1-	*All* P	*All* R	*All* F1
Phrase Replacement	CE	✗	70	56	62	78	87	82	75	76	75
		✓	61	81	69	87	71	78	78	75	75
	Margin	✓	70	76	73	86	82	84	80	80	80
Subtree Deletion	CE	✗	81	50	62	77	93	85	78	78	77
		✓	64	84	72	89	74	81	80	77	78
	Margin	✓	70	83	76	90	81	85	83	82	82
Lexical Substitution	CE	✗	65	53	57	76	84	80	72	73	72
		✓	55	81	66	86	64	73	75	70	71
	Margin	✓	57	75	65	83	70	76	74	72	72
Balanced	CE	✗	76	42	54	74	93	83	75	75	73
		✓	73	73	73	85	85	85	81	81	81
	Margin	✓	76	73	75	85	87	86	82	82	82
Concatenation	CE	✗	62	32	42	70	89	79	67	69	66
		✓	73	55	63	78	89	83	76	77	76
	Margin	✓	84	59	70	81	94	87	82	82	81
Divergence Ranking	Margin	✓	82	72	**77**	86	91	**88**	84	85	**84**
LASER baseline			38	58	46	68	48	57	57	52	53

Table 3: Intrinsic evaluation of Divergent mBERT and its ablation variants on the REFRESD dataset. We report Precision (P), Recall (R), and F1 for the equivalent (+) and divergent (-) classes separately, as well as for both classes (*All*). *Divergence Ranking* yields the best F1 scores across the board.



(a) LASER (b) *Subtree Deletion* (c) *Divergence Ranking*

Figure 3: Score distributions assigned by different models to sentence-pairs of REFRESD. *Divergence Ranking* scores for the "Some meaning difference" class are correctly skewed more toward negative values.

Contrastive Samples With the CE loss, independent contrastive samples improve over randomly sampled synthetic instances overall (+8.7 F1+ points on average), at the cost of a smaller drop for the divergent class (−5.3 F1- points) for models trained on a single type of divergence. Using the margin loss helps models recover from this drop.

Divergence Types All types improve over the LASER baseline. When using a single divergence type, *Subtree Deletion* performs best, even matching the overall F1 score of a system trained on all types of divergences (*Balanced Sampling*). Training on the *Concatenation* of all divergence types yields poor performance. We suspect that the model is overwhelmed by negative instances at training time, which biases it toward predicting the divergent class too often and hurting F1+ score for the equivalent class.

Divergence Ranking How does divergence ranking improve predictions? Figure 3 shows model score distributions for the 3 classes annotated in REFRESD. *Divergence Ranking* particularly improves divergence predictions for the "Some meaning difference" class: the score distribution for this class is more skewed toward negative values than when training on contrastive *Subtree Deletion* samples.

		Union			Pair-wise Union			Intersection		
Model	Multi-task	F1-DIV	F1-EQ	F1-Mul	F1-DIV	F1-EQ	F1-Mul	F1-DIV	F1-EQ	F1-Mul
Random Baseline		0.21	0.62	0.13	0.33	0.59	0.20	0.21	0.62	0.13
Token-only		0.39	0.77	0.30	0.46	**0.88**	0.41	0.46	**0.92**	0.42
Balanced	✓	0.41	0.77	0.32	0.46	0.87	0.40	0.43	0.91	0.40
Concatenation	✓	0.41	**0.78**	0.32	0.48	**0.88**	0.42	0.46	**0.92**	0.42
Divergence Ranking	✓	**0.45**	**0.78**	**0.35**	**0.51**	**0.88**	**0.45**	**0.49**	**0.92**	**0.45**

Table 4: Evaluation of different models on the token-level prediction task for the **"Some meaning difference"** class of REFRESD. *Divergence Ranking* yields the best results across the board.

7 Finer-Grained Divergence Detection

While we cast divergence detection as binary classification in Section 6, human judges separated divergent samples into "Unrelated" and "Some meaning difference" classes in the REFRESD dataset. Can we predict this distinction automatically? In the absence of annotated training data, we cannot cast this problem as a 3-way classification, since it is not clear how the synthetic divergence types map to the 3 classes of interest. Instead, we test the hypothesis that token-level divergence predictions can help discriminate between divergence granularities at the sentence-level, inspired by humans' use of rationales to ground sentence-level judgments.

7.1 Experimental Conditions

Models We fine-tune the **multi-task** MBERT model that makes token and sentence predictions jointly, as described in Section 3.4. We contrast against a sequence labeling MBERT model trained independently with the CE loss (**Token-only**). Finally, we run a **random baseline** where each token is labeled EQ or DIV uniformly at random.

Training Data We tag tokens edited when generating synthetic divergences as DIV (e.g., highlighted tokens in Table 1), and others as EQ. Since edit operations are made on the English side, we tag aligned French tokens using the Berkeley aligner.

Evaluation We expect token-level annotations in REFRESD to be noisy since they are produced as rationales for sentence-level rather than token-level tags. We, therefore, consider three methods to aggregate rationales into token labels: a token is labeled as DIV if it is highlighted by at least one (**Union**), two (**Pair-wise Union**), or all three annotators (**Intersection**). We report F1 on the DIV and EQ class, and F1-Mul as their product for each of the three label aggregation methods.

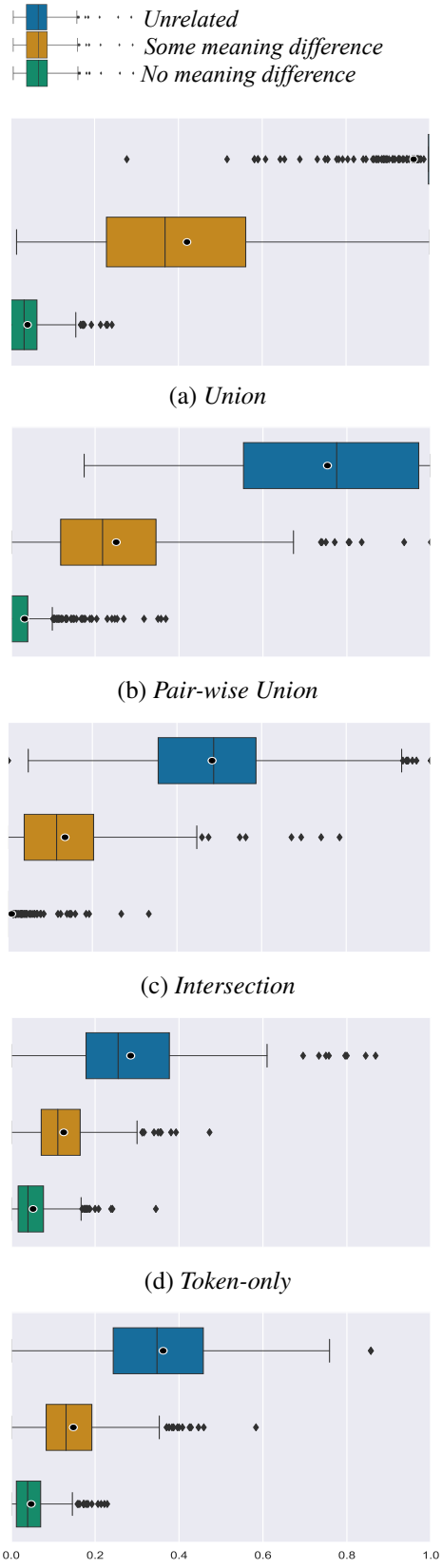
7.2 Results

Token Labeling We evaluate token labeling on REFRESD samples from the "Some meaning difference" class, where we expect the more subtle differences in meaning to be found, and the token-level annotation to be most challenging (Table 4). Examples of Divergent MBERT's token-level predictions are given in A.6. The *Token-only* model outperforms the *Random Baseline* across all metrics, showing the benefits of training even with noisy token labels derived from rationales. *Multi-task* training further improves over *Token-only* predictions for almost all different metrics. *Divergence Ranking* of contrastive instances yields the best results across the board. Also, on the auxiliary sentence-level task, the *Multi-task* model matches the F1 as the standalone *Divergence Ranking* model.

From Token to Sentence Predictions We compute the % of DIV predictions within a sentence-pair. The multi-task model makes more DIV predictions for the divergent classes as its % distribution is more skewed towards greater values (Figure 4 (d) vs. (e)). We then show that the % of DIV predictions of the *Divergence Ranking* model can be used as an indicator for distinguishing between divergences of different granularity: intuitively, a sentence pair with more DIV tokens should map to a coarse-grained divergence at a sentence-level. Table 5 shows that thresholding the % of DIV tokens could be an effective discrimination strategy, which we will explore further in future work.

	UN			SD			
DIV %	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1-all**
10	48	**97**	64	66	51	57	59
20	69	84	**76**	**83**	79	81	80
30	82	63	71	81	**85**	**83**	**81**
40	**94**	35	51	73	84	78	75

Table 5: "Some meaning difference" (SD) vs. "Unrelated" (UN) classification based on % of DIV labels.



(a) *Union*

(b) *Pair-wise Union*

(c) *Intersection*

(d) *Token-only*

(e) *Divergence Ranking*

Figure 4: Percentage distributions of DIV tokens in REFRESD and DIV token predictions of two models: *Divergence Ranking* makes more DIV predictions compared to the *Token-only* model, enabling a better distinction between the divergent classes.

8 Related Work

Our work is closely related to but distinct from the Semantic Textual Similarity (STS) task that measures the **degree of equivalence** in the underlying semantics of paired snippets of text (Agirre et al., 2016; Cer et al., 2017). Most commonly, state-of-the-art models address the STS task via interaction models that use alignment mechanisms to integrate word-level interactions in their final predictions (He and Lin, 2016; Parikh et al., 2016) or via learning vector representations of sentences that are then compared using distance-based measures (Nie and Bansal, 2017; Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019; Yang et al., 2019).

9 Conclusion

We show that explicitly considering diverse semantic divergence types benefits both the annotation and prediction of divergences between texts in different languages.

We contribute REFRESD, a new dataset of WikiMatrix sentences-pairs in English and French, annotated with semantic divergence classes and token-level rationales that justify the sentence level annotation. 64% of samples are annotated as divergent, and 40% of samples contain fine-grained meaning divergences, confirming that divergences are too frequent to ignore even in parallel corpora. We show that these divergences can be detected by a MBERT model fine-tuned without annotated samples, by learning to rank synthetic divergences of varying granularity.

Inspired by the rationale-based annotation process, we show that predicting token-level and sentence-level divergences jointly is a promising direction for further distinguishing between coarser and finer-grained divergences.

Acknowledgements

We thank the anonymous reviewers and the CLIP lab at UMD for helpful comments. This material is based upon work supported by the National Science Foundation under Award No. 1750695. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016. SemEval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the Wikipedia Language Gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1075–1084, Austin, Texas, USA. ACM.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Rémi Cardon and Natalia Grabar. 2020. Reducing the search space for parallel sentences in comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 44–48, Marseille, France. European Language Resources Association.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

K. Durán, J. Rodríguez, and M. Bravo. 2014. Similarity of sentences through comparison of syntactic trees with pairs of similar words. In *2014 11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–6.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, Geneva, Switzerland. COLING.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *ArXiv*, abs/1803.09010.

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 Transla- tion in the Americas*.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 937–948, San Diego, California. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2019a. Multi-task stack propagation for neural quality estimation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 2019b. QE BERT: Bilingual BERT using multi-task learning for neural quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 85–89, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark. Association for Computational Linguistics.

Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 618–626, New York, NY, USA. ACM.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Cross-lingual Discourse Relation Analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 577–587, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823, Florence, Italy. Association for Computational Linguistics.

Jun Lu, Xiaoyu Lv, Yangbin Shi, and Boxing Chen. 2018. Alibaba submission to the WMT18 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 917–922, Belgium, Brussels. Association for Computational Linguistics.

Paolo Massa and Federico Scrinzi. 2012. Manypedia: Comparing Language Points of View of Wikipedia Communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 21:1–21:9, Linz, Austria. ACM.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta : journal des traducteurs / Meta: Translators' Journal*, 47(4):498–512.

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

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45, Copenhagen, Denmark. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

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

MinhQuang Pham, Josep Crego, Jean Senellart, and François Yvon. 2018. Fixing translation divergences in parallel corpora for neural MT. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973, Brussels, Belgium. Association for Computational Linguistics.

Jipeng Qiang, Yifan Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2019. A simple bert-based approach for lexical simplification. *ArXiv*, abs/1907.06226.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *ArXiv*, abs/1907.05791.

Margita Šoštarić, Christian Hardmeier, and Sara Stymne. 2018. Discourse-related language contrasts in English-Croatian human and machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 36–48, Brussels, Belgium. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing Wikipedia Across Languages via Recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 975–985, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Ziyi Yang, Chenguang Zhu, and Weizhu Chen. 2019. Parameter-free sentence embedding via orthogonal basis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 638–648, Hong Kong, China. Association for Computational Linguistics.

Ching-Man Au Yeung, Kevin Duh, and Masaaki Nagata. 2011. Providing Cross-Lingual Editing Assistance to Wikipedia Editors. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 377–389, Tokyo, Japan. Springer-Verlag.

Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL HLT 2007; Proceedings of the Main Conference*, pages 260–267.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with

translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat. 2019. Towards recognizing phrase translation processes: Experiments on english-french. *ArXiv*, abs/1904.12213.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

A Appendices

A.1 Implementation Details

Training setup We employ the Adam optimizer with initial learning rate $\eta = 2\text{e}{-5}$, fine-tune for at most 5 epochs, and use early-stopping to select the best model. We use a batch size of 32 for experiments that do not use contrastive training and a batch size of 16 for those using contrastive training to establish a fair comparison.

Model setup All of our models are based on the "Multilingual BERT-base model" consisting of: 12-layers, 768-hidden size, 12-heads and 110M parameters.

Average Runtime & Computing Infrastructure Each experiment is run on a single GeForce GTX 1080 GPU. For experiments run on either a single type of divergence (e.g., *Subtree Deletion*) or using *Balanced* sampling, the average duration time is $\sim$ 0.4 hours. For *Divergence Ranking* and *Concatenation*, sampling methods, training takes $\sim$ 2 hours to complete.

Hyperparameter search on margin We perform a grid search on the margin parameter for each experiment that employs contrastive training. We experiment with values $\{3, 4, 5, 6, 7, 8\}$ and pick the one corresponding to the best Weighted-F1 score on a synthetic development set. Table 6 shows mean and variance results on both the development and the REFRESD dataset for different ξ values. In general, we observe that our model's performance on REFRESD is not sensitive to the margin's choice, as reflected by the small variances on the REFRESD Weighted-F1.

A.2 Very Deep Pair-Wise Interaction baseline

We compare against the Very Deep Pair-Wise Interaction (VDPWI) model repurposed by Vyas et al. (2018) to identify cross-lingual semantic divergence vs. equivalence. We fine-tune MBERT models on coarsely-defined semantic synthetic divergent pairs, similarly to the authors. We report results on two crowdsourced datasets, consisting of equivalence vs. divergence labels for 300 sentence-pairs, drawn from the noisy OpenSubtitles and CommonCrawl corpora. The two evaluation datasets are available at: https://github.com/yogarshi/SemDiverge/tree/master/dataset.

Synthetic	**Dev**	**REFRESD**	ξ*	**Dev***
Phrase replacement	$91.83_{\pm 1.14}$	$78.60_{\pm 1.84}$	7	93
Subtree Deletion	$93.67_{\pm 2.22}$	$82.67_{\pm 0.22}$	8	95
Lexical Substitution	$91.50_{\pm 0.25}$	$70.50_{\pm 1.58}$	5	92
Balanced	$87.67_{\pm 0.56}$	$80.33_{\pm 0.56}$	5	88
Concatenation	$89.03_{\pm 0.50}$	$79.51_{\pm 0.67}$	5	90
Divergence Ranking	$77.80_{\pm 1.36}$	$83.67_{\pm 0.22}$	5	79

Table 6: Average results of Divergent MBERT as a function of the number of hyperparameter trials for the margin value (ξ). The first row corresponds to the sampling method used for creating synthetic contrastive training examples. The second and third rows correspond to the mean/variance of Weighted-F1 results, measured on the development and the REFRESD dataset, respectively. The fourth row describes the best value of the margin hyperparameter (ξ*) for each experiment, while the last row denotes the corresponding Weighted-F1 score on the development set.

	OpenSubtitles			CommonCrawl		
Method	**F1-**	**F1+**	**F1**	**F1-**	**F1+**	**F1**
Vyas et al. (2018)	78	72	77	85	73	80
MBERT	**81**	**76**	**79**	**87**	**76**	**83**

Table 7: Performance comparison between MBERT and VDPWI trained on coarsely-generated semantic divergences. We report F1 overall results (F1) and F1+/F1- scores for the two classes, on the crowdsourced OpenSubtitles and CommonCrawl datasets.

Table 7 presents results on the OpenSubtitles and CommonCrawl testbeds. We observe that MBERT trained on similarly defined coarse divergences performs better than cross-lingual VDPWI.

A.3 REFRESD: Annotation Guidelines

Below we include the annotation guidelines given to participants:

"You are asked to compare the meaning of English and French text excerpts. You will be presented with one pair of texts at a time (about a sentence in English and a sentence in French). For each pair, you are asked to do the following:

1 Read the two sentences carefully. Since the sentences are provided out of context, your understanding of content should only rely on the information available in the sentences. There is no need to guess what additional information might be available in the documents the excerpts come from.

2 Highlight the text spans that convey different meaning in the two sentences. After highlighting a span of text, you will be asked to further characterize it as:

ADDED *the highlighted span corresponds to a piece of information that* ***does not exist*** *in the other sentence*
CHANGED *the highlighted span corresponds to a piece of information that exists in the other sentence, but* ***their meaning is not the exact same***
OTHER *none of the above holds*

You can highlight as many spans as needed. You can **optionally** provide an explanation for your assessment in the text form under the Notes section (e.g., literal translation of idiom)

3 Compare the meaning of the two sentences by picking one of the three classes:

UNRELATED *The two sentences are completely unrelated or have a few words in common but convey unrelated information about them*
SOME MEANING DIFFERENCE *The two sentences convey* ***mostly the same information, except differences for some details or nuances*** *(e.g., some information is added and/or missing on either or both sides; some English words have a more general or specific translation in French)*
NO MEANING DIFFERENCE *The two sentences have the* ***exact same meaning****"*

A.4 Annotation Procedures

We run 8 online annotation sessions. Each session consists of 120 instances, annotated by 3 participants, and lasts about 2 hours. Participants are allowed to take breaks during the process. Participants are rewarded with Amazon gift cards at a rate of $2 per 10 examples, with bonuses of $5 and $10 for completing the first and additional sessions, respectively.

A.5 Annotated examples in REFRESD

Table 8 includes examples of annotated instances drawn from REFRESD, corresponding to different levels of inter-annotator agreement.

A.6 Token predictions of Divergent mBERT

Table 9 shows randomly selected instances from REFRESD along with token tags predicted by our best performing system (*Divergence Ranking*).

A.7 Results on synthetic development sets

Tables 10 and 11 report results on development sets for each experiment included in Tables 3 and 4, respectively.

	No meaning difference with *high* sentence-level agreement and *high* span overlap (n=3)
EN	The plan was revised in 1916 to concentrate the main US naval fleet in New England, and from there defend the US from the German navy.
FR	Le plan fut révisé en 1916 pour concentrer le gros de la flotte navale américaine en Nouvelle-Angleterre, et à partir de là, défendre les États-Unis contre la marine allemande.
	Some meaning difference with *high* sentence-level agreement and *high* span overlap (n=3)
EN	After an intermediate period during which Stefano Piani edited the stories, in 2004 a major rework of the series went through.
FR	Après une période intermédiaire pendant laquelle Stefano Piani édita les histoires, une refonte majeure de la série fut faite en 2004 en réponse à une baisse notable des ventes.
	Unrelated with *high* sentence-level agreement and *high* span overlap (n=3)
EN	To reduce vibration, all helicopters have rotor adjustments for height and weight.
FR	En vol, le régime du compresseur Tous les compresseurs ont un taux de compression lié à la vitesse de rotation et au nombre d'étages.
	No meaning difference with *high* sentence-level agreement and *high* span overlap (n=3)
EN	One can see two sunflowers on the main façade and three smaller ones on the first floor above ground just above the entrance arcade.
FR	On remarquera deux tournesols sur la façade principale et trois plus petits au premier étage au-dessus des arcades d'entrée.
	Some meaning difference with *high* sentence-level agreement and *low* span overlap (n=3)
EN	On November 10, 2014, CTV ordered a fourth season of Saving Hope that consisted of eighteen episodes, and premiered on September 24.
FR	Le 10 novembre 2014, CTV a renouvelé la série pour une quatrième saison de 18 épisodes diffusée depuis le 24 septembre 2015.
	Unrelated with *high* sentence-level agreement and *low* span overlap (n=3)
EN	He talks about Jay Gatsby, the most hopeful man he had ever met .
FR	Il côtoie notamment Giuseppe Meazza qui dira de lui Il fut le joueur le plus fantastique que j'aie eu l'occasion de voir.
	No meaning difference with *moderate* sentence-level agreement (n=2)
EN	Nine of these revised BB LMs were built by Ferrari in 1979, while a further refined series of sixteen were built from 1980 to 1982.
FR	Neuf de ces BB LM révisées furent construites par Ferrari en 1979, tandis qu'une série de seize autres furent construite entre 1980 et 1982.
	Some meaning difference with *moderate* sentence-level agreement (n=2)
EN	From 1479, the Counts of Foix became Kings of Navarre and the last of them , made Henri IV of France, annexed his Pyrrenean lands to France.
FR	À partir de 1479, le comte de Foix devient roi de Navarre et le dernier d'entre eux, devenu Henri IV, roi de France en 1607, annexe ses terres pyrénéennes à la France.
	Unrelated difference with *moderate* sentence-level agreement (n=2)
EN	The operating principle was the same as that used in the Model 07/12 Schwarzlose machine gun used by Austria-Hungary during the First World War.
FR	Le Skoda 100 mm modèle 1916 était un obusier de montagne utilisé par l'Autriche-Hongrie pendant la Première Guerre mondiale.

Table 8: REFRESD examples, corresponding to different levels of agreement between annotators. n denotes the number of annotators who voted for the sentence-level majority class; disagreements span closely related classes.

EN	He joined the Munich State Opera in 1954, where he created the role of Johannes Kepler in Hindemith's Die Harmonie der Welt (1957). He joined the Munich State Opera in 1954, where he created the role of Johannes Kepler in Hindemith's Die Harmonie der Welt (1957).
FR	Il crée à Munich, le rôle de Johannes Kepler dans Die Harmonie der Welt de Paul Hindemith en 1957. Il crée à Munich, le rôle de Johannes Kepler dans Die Harmonie der Welt de Paul Hindemith en 1957.
EN	He experimented with silk vests resembling medieval gambesons, which used 18 to 30 layers of silk fabric to protect the wearers from penetration. He experimented with silk vests resembling medieval gambesons, which used 18 to 30 layers of silk fabric to protect the wearers from penetration.
FR	Ils ressemblaient aux jaques, vêtements matelassés médiévaux constitués de 18 à 30 couches de vêtements afin d'offrir une protection maximale contre les flèches. Ils ressemblaient aux jaques, vêtements matelassés médiévaux constitués de 18 à 30 couches de vêtements afin d'offrir une protection maximale contre les flèches.
EN	Even though this made Armenia a client kingdom, various contemporary Roman sources thought that Nero had de facto ceded Armenia to the Parthian Empire. Even though this made Armenia a client kingdom, various contemporary Roman sources thought that Nero had de facto ceded Armenia to the Parthian Empire.
FR	Plusieurs sources romaines contemporaines n'en ont pas moins considéré que Néron a ainsi de facto cédé l'Arménie aux Parthes. Plusieurs sources romaines contemporaines n'en ont pas moins considéré que Néron a ainsi de facto cédé l'Arménie aux Parthes.
EN	The Photo League was a cooperative of photographers in New York who banded together around a range of common social and creative causes. The Photo League was a cooperative of photographers in New York who banded together around a range of common social and creative causes.
FR	La Photo League était un groupement de photographes amateurs et professionnels réuni à New York autour d'objectifs communs de nature sociale et créative. La Photo League était un groupement de photographes amateurs et professionnels réuni à New York autour d'objectifs communs de nature sociale et créative.
EN	She made a courtesy call to the Hawaiian Islands at the end of the year and proceeded thence to Puget Sound where she arrived on 2 February 1852. She made a courtesy call to the Hawaiian Islands at the end of the year and proceeded thence to Puget Sound where she arrived on 2 February 1852.
FR	Il fait une escale aux îles Hawaï à la fin de l'année, au Puget Sound, le 2 février 1852. Il fait une escale aux îles Hawaï à la fin de l'année, au Puget Sound, le 2 février 1852.
EN	Recognizing Nishikaichi and his plane as Japanese, Kaleohano thought it prudent to relieve the pilot of his pistol and papers before the dazed airman could react. Recognizing Nishikaichi and his plane as Japanese, Kaleohano thought it prudent to relieve the pilot of his pistol and papers before the dazed airman could react .
FR	Reconnaissant Nishikaichi et son avion comme étant japonais, Kaleohano pensa qu'il serait prudent de confisquer au pilote son pistolet et ses documents. Reconnaissant Nishikaichi et son avion comme étant japonais, Kaleohano pensa qu 'il serait prudent de confisquer au pilote son pistolet et ses documents.
EN	At the same time , the mortality rate increased slightly from 8.9 per 1,000 inhabitants in 1981 to 9.6 per 1,000 inhabitants in 2003. At the same time, the mortality rate increased slightly from 8.9 per 1,000 inhabitants in 1981 to 9.6 per 1,000 inhabitants in 2003 .
FR	Le taux de mortalité est quant à lui passé de 11,8 % sur la période 1968-1975 à 9,1 % sur la période 1999-2009 . Le taux de mortalité est quant à lui passé de 11,8 % sur la période 1968-1975 à 9,1 % sur la période 1999-2009.
EN	They called for a state convention on September 17 in Columbia to nominate a statewide ticket. They called for a state convention on September 17 in Columbia to nominate a statewide ticket.
FR	Un décret de la Convention du 28 avril 1794 ordonna que son nom fût inscrit sur une colonne de marbre au Panthéon. Un décret de la Convention du 28 avril 1794 ordonna que son nom fût inscrit sur une colonne de marbre au Panthéon.
EN	His plants are still in the apartment and the two take all of the plants with them back to their place. His plants are still in the apartment and the two take all of the plants with them back to their place.
FR	Il reste donc chez lui et les deux sœurs s'occupent du show toutes seules. Il reste donc chez lui et les deux sœurs s'occupent du show toutes seules.

Table 9: REFRESD examples, along with Divergent mBERT's predictions. Tokens highlighted with green color correspond to DIV predictions of Divergent mBERT (second sentence). Tokens highlighted with red colors correspond to gold-standard labels of divergence provided by annotators (first sentence). The red color intensity denotes the degree of agreement across three annotators (darker color denotes higher agreement).

Divergent	Loss	Contrastive	*Equivalents*			*Divergents*			*All*		
			P+	**R+**	**F1+**	**P-**	**R-**	**F1-**	**P**	**R**	**F1**
Phrase Replacement	CE	✗	92	97	94	96	92	94	94	94	94
		✓	92	97	94	97	91	94	94	94	94
	Margin	✓	91	95	93	95	91	93	93	93	93
Subtree Deletion	CE	✗	93	97	95	97	93	95	95	95	95
		✓	94	97	96	97	94	96	96	96	96
	Margin	✓	93	97	95	97	93	95	95	95	95
Lexical Substitution	CE	✗	93	94	94	94	93	93	94	94	93
		✓	95	93	94	94	95	94	94	94	94
	Margin	✓	91	94	93	94	91	92	92	92	92
Balanced	CE	✗	90	96	92	95	89	92	92	92	92
		✓	90	94	92	94	90	92	92	92	92
	Margin	✓	85	93	89	92	84	88	89	88	88
Concatenation	CE	✗	92	90	91	90	92	91	91	91	91
		✓	82	89	86	97	95	96	94	94	94
	Margin	✓	89	92	90	91	88	90	90	90	90
Divergence Ranking	Margin	✓	72	96	82	94	63	75	83	79	79

Table 10: Evaluation on **synthetic development sets**. We report Precision (P), Recall (R), and F1 for the equivalent (+) and divergent (-) classes separately and both classes (All). Each model uses a development set that includes divergent types used during training. *Divergence Ranking* yields lower performance on the synthetic development set than REFRESD, reflecting the mismatch between the nature of synthetics samples vs. divergences in REFRESD.

Model	Multi-task	F1-EQ	F1-DIV	F1-Mul
Token-only		99	88	87
Balanced	✓	98	71	70
Concatenation	✓	98	71	70
Divergence Ranking	✓	98	75	74

Table 11: Evaluation of token tagging models on **synthetic test sets**. We report Precision (P), Recall (R), and F1 scores for each class. F1-Mul corresponds to the product of individual F1 scores. The model's performance on synthetic test sets is always better than the one reported on REFRESD, reflecting the mismatch between the noisy training samples and the real divergences found in REFRESD.