Probing Multilingual Language Models for Discourse

Murathan Kurfalı
Linguistics Department
Stockholm University
Stockholm, Sweden
`murathan.kurfali@ling.su.se`

Robert Östling
Linguistics Department
Stockholm University
Stockholm, Sweden
`robert@ling.su.se`

Abstract

Pre-trained multilingual language models have become an important building block in multilingual natural language processing. In the present paper, we investigate a range of such models to find out how well they transfer discourse-level knowledge across languages. This is done with a systematic evaluation on a broader set of discourse-level tasks than has been previously been assembled. We find that the XLM-RoBERTa family of models consistently show the best performance, by simultaneously being good monolingual models and degrading relatively little in a zero-shot setting. Our results also indicate that model distillation may hurt the ability of cross-lingual transfer of sentence representations, while language dissimilarity at most has a modest effect. We hope that our test suite, covering 5 tasks with a total of 22 languages in 10 distinct families, will serve as a useful evaluation platform for multilingual performance at and beyond the sentence level.

1 Introduction

Large-scale pre-trained neural language models have become immensely popular in the natural language processing (NLP) community in recent years (Devlin et al., 2019; Peters et al., 2018). When used as contextual sentence encoders, these models have led to remarkable improvements in performance for a wide range of downstream tasks (Qiu et al., 2020). In addition, multilingual versions of these models (Devlin et al., 2019; Conneau and Lample, 2019) have been successful in transferring knowledge across languages by providing language-independent sentence encodings.

The general usefulness of pre-trained language models has been convincingly demonstrated thanks to persistent creation and application of evaluation datasets by the NLP community. Discourse-level analysis is particularly interesting to study, given that many of the currently available models are trained with relatively short contexts such as pairs of adjacent sentences.

Wang et al. (2019) use a diverse set of natural language understanding (NLU) tasks to investigate the generality of the sentence representations produced by different language models. Hu et al. (2020) use a broader set of tasks from across the NLP field to investigate the ability of multilingual models to transfer various types of knowledge across language boundaries.

Our goal in this paper is to systematically evaluate the multilingual performance on NLU tasks, particularly at the discourse level. This combines two of the most challenging aspects of representation learning: multilinguality and discourse-level analysis. A few datasets have been used for this purpose before, most prominently the XNLI evaluation set (Conneau et al., 2018) for Natural Language Inference (NLI), and recently also XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020) for Question Answering (QA). We substantially increase the breadth of our evaluation by adding three additional tasks:

1. Penn Discourse TreeBank (PDTB)-style implicit discourse relation classification on annotated TED talk subtitles in seven languages (Section 3.1.1)

2. Rhetorical Structure Theory (RST)-style discourse relation classification with a custom set consisting of treebanks in six non-English languages (Section 3.1.2)

3. Stance detection with a custom dataset in five languages (Section 3.1.3)

We investigate the cross-lingual generalization capabilities of seven multilingual sentence encoders with considerably varying model sizes

through their cross-lingual zero-shot performance[1] which, in this context, refers to the evaluation scheme where sentence encoders are tested on the languages that they are not exposed to during training. The complied test suite consists of five tasks, covering 22 different languages in total.

We specifically focus on zero-shot transfer scenario where a sufficient amount of annotated data to fine-tune a pre-trained language model is assumed to be available only for one language. We believe that this is the most realistic scenario for a great number of languages; therefore, zero-shot performance is the most direct way of assessing cross-lingual usefulness in a large scale.

Our contributions are as follows: (i) we provide a detailed analysis of a wide range of sentence encoders on large number of probing tasks, several of which have not previously been used with multilingual sentence encoders despite their relevancy, (ii) we provide suitably pre-processed versions of these datasets to be used as a multilingual benchmark for future work with strong baselines provided by our evaluation, (iii) we show that the zero-shot performance on discourse level tasks are not correlated with any kind of language similarity and hard to predict, (iv) we show that knowledge distillation may selectively destroy multilingual transfer ability in a way that harms zero-shot transfer, but is not visible during evaluations where the models are trained and evaluated with the same language.

2 Background

The standard way of training a multilingual language model is through a large non-parallel multilingual corpora, e.g. Wikipedia articles, where the models are not provided with any explicit mapping across languages which renders cross-lingual performance of such models puzzling. Pires et al. (2019) and Wu and Dredze (2019) are the earliest studies to explore that puzzle by trying to uncover the factors that give multilingual BERT (henceforth, mBERT) its cross-lingual capabilities. Pires et al. (2019) perform a number of probing tasks and hypothesize that the shared sentence pieces across languages gives mBERT its generalization ability by forcing other pieces to be mapped into the same space. Similarly, Wu and Dredze (2019)

evaluate the performance of mBERT in five tasks and report that while mBERT shows a strong zero-shot performance, it also retains language-specific information in each layer.

Chen et al. (2019a) proposes a benchmark to evaluate sentence encoders specifically on discourse level tasks. The proposed benchmark consists of discourse relation classification and a number of custom tasks such as finding the correct position of a randomly moved sentence in a paragraph or determining if a given paragraph is coherent or not. The benchmark is confined to English, hence, only targets monolingual English models.

Two very recent studies, XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020), constitute the first studies on the cross-lingual generalization abilities of pre-trained language models via their zero-shot performance. The tasks in both studies largely overlap, where XTREME serves as cross-lingual benchmark consisting of well-known datasets, e.g. XNLI, XQuAD. On the other hand, while covering the most of XTREME tasks[2], XGLUE offers new datasets which either focus on the relation between a pair of inputs, such as web page–query matching, or on text generation via question/news title generation. In addition to the mBERT and certain XLM and XLM-R versions, XTREME includes MMTE (Arivazhagan et al., 2019) whereas XGLUE evaluates Unicoder (Huang et al., 2019) among its baselines.

3 Cross-lingual Discourse-level Evaluation

In discourse research, sentences/clauses are not understood in isolation but in relation to one another. The semantic interactions between these units are usually regarded as the backbone of coherence in various prominent discourse theories including that underlying the Penn Discourse TreeBank (PDTB) (Prasad et al., 2007), and Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) used in the RST Discourse Treebank (Carlson and Marcu, 2001). Modelling such interactions requires an understanding that is beyond sentence-level and, from this point-of-view, determining any kind of relation between sentences/clauses can be associated with discourse.

Although paraphrase detection or natural language inference may not strike as discourse-level tasks at first glance, they both deal with semantic

---

[1] In the remainder of the paper, *cross-lingual zero-shot performance* is simply referred as *zero-shot performance* for brevity. Similarly, source language performance denotes the performance of the respective model on the test set of the training language.

[2] Except parallel sentence retrieval tasks.

relations between sentences. Tonelli and Cabrio (2012) show that textual entailment is, in fact, a sub-class of *Restatement* relations of the PDTB framework whereas Nie et al. (2019) report an increase in discourse relation classification accuracy when NLI is used as the intermediate fine-tuning task. In a similar vein, a stance against a judgement, *Favor* or *Against*, can be seen as *CONTINGENCY: Cause: reason* and *COMPARISON: Contrast* in PDTB; *Explanation* and *Antithesis* in RST, respectively.

Therefore, these NLU tasks can be seen as special subsets of discourse relation classification; only a model with a good understanding beyond individual sentences can be expected to solve these tasks. Finally, since question answering requires an understanding on discourse level in order to be solved, so we also believe classifying this as a discourse-level task should be uncontroversial.

3.1 Tasks & Datasets

In this section, we present our task suite and the datasets used for training and zero-shot evaluation. For the sake of clarity, we name each task after the dataset used for training.

3.1.1 Implicit Discourse Relation Classification (PDTB)

Implicit discourse relations hold between adjacent sentence pairs but are not explicitly signaled with a connective such as *because*, *however*. Implicit discourse relation classification is the task of determining the sense conveyed by these adjacent sentences, which can be easily inferred by readers. Classifying implicit relations constitutes the most challenging step of shallow discourse parsing (Xue et al., 2016).

The training is performed on PDTB3 (Webber et al., 2016) where sections 2–20, 0–1 are used for training and development respectively. The zero-shot evaluation is performed on the TED-MDB corpus (Zeyrek et al., 2019)[3], which is a PDTB-style annotated parallel corpus consisting of 6 TED talk transcripts, and the recent Chinese annotation effort on TED talk transcripts that however are mostly not parallel to TED-MDB (Long et al., 2020). Due to the small size of the test sets, we confine ourselves to the top-level senses: *Contingency, Comparison, Expansion, Temporal* which is also the most common setting for this task. Despite the limited size of TED-MDB, zero-shot transfer is possible and

yields meaningful results as shown in (Kurfalı and Östling, 2019). In total, seven languages are evaluated in this task: English, German, Lithuanian[4], Portuguese, Polish, Russian and Chinese.

3.1.2 Rhetorical Relation Classification (RST)

Rhetorical relations are just another name for discourse relations but this term is most commonly associated with Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). Similar to PDTB's discourse relations, rhetorical relations also denote links between discourse units, but are considerably different from the former. The difference largely stems from the take of the respective theories on the structure of the discourse. RST conceives discourse as one connected tree-shaped structure assuming hierarchical relations among the discourse relations. On the other hand, PDTB does not make any claims regarding the structure of the discourse and annotates discourse relations only in a local context (i.e. adjacent clauses/sentences) without assuming any relation on higher levels. Hence, evaluation on RST and PDTB relations can be seen as complementary to each other as the former focuses on both global and local discourse structure whereas PDTB focuses only on local structure.

We use English RST-DT (Carlson and Marcu, 2001) for training where a randomly selected 35 documents are reserved for development. However, unlike PDTB, there is not any compact parallel RST corpus; RST annotations across languages usually differ from each other in several ways. Therefore, we follow Braud et al. (2017) and create a custom multilingual corpus for the zero-shot experiments which consists of the following languages: Basque (Iruskieta et al., 2013), Brazilian Portuguese (Cardoso et al., 2011; Collovini et al., 2007; Pardo and Seno, 2005), Chinese (Cao et al., 2018), German (Stede, 2004), Spanish (Da Cunha et al., 2011), Russian (Pisarevskaya et al., 2017). We perform a normalization step on each treebank which includes binarization of non-binary trees and mapping all relations to 18 coarse grained classes described in (Carlson and Marcu, 2001). The normalization step is performed via the pre-processing scripts of (Braud et al., 2017). Due to memory constraints, we limit the sequence lengths to 384. Hence, we only keep those relations where the first discourse unit is shorter than 150 words so that both units can

---

[3]https://github.com/MurathanKurfali/Ted-MDB-Annotation

[4]Lithuanian is the latest addition to the Ted-MDB corpus, as documented in (Oleskeviciene et al., 2018).

be equally represented which lead to omission of only 5% of all non-English relations.

3.1.3 Stance Detection (X-Stance)

The stance detection is task of determining the attitude expressed in a text towards a target claim. For experiments, we mainly use the X-stance corpus which consists of 60K answers to 150 questions concerning politics in German, Italian and French (Vamvas and Sennrich, 2020). Unlike other tasks, we select German as the training language for stance detection as it is the largest language in X-Stance. Following the official split, we use the German instances in the training and development sets during fine-tuning and non-German instances in the test set for evaluation. Furthermore, we enrich the scope of our zero-shot evaluation by two additional dataset, one in English (Chen et al., 2019b) and other one in Chinese (Yuan et al., 2019), which also consist of stance annotated claim–answer pairs, despite in different domains.

3.1.4 Natural Language Inference (XNLI)

Natural language inference (NLI) is the task of determining whether a premise sentence entails, contradicts or is neutral to a hypothesis sentence. MultiNLI and the mismatched part of the development data (Williams et al., 2018) are used for training and validation, respectively. The evaluation is performed on the test sets of the XNLI (Conneau et al., 2018) corpus which covers the following 14 languages in addition to English: French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu.

3.1.5 Question Answering (XQuAD)

Question answering is the task of identifying span in a paragraph which answers to a question. We use the SQuAD v1.1 (Rajpurkar et al., 2016) for training. We evaluate the models on the popular XQuAD dataset which contains the translation of SQuAD v1.1 development set into ten languages (Artetxe et al., 2020): Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi.

3.2 Languages

The proposed task suite covers the following 22 languages representing 10 language families: Indo-European (Bulgarian *bg*, German *de*, Greek *el*, English *en*, Spanish *es*, French *fr*, Hindi *hi*, Italian

it, Lithuanian *lt*, Polish *pl*, Portuguese *pt*, Russian *ru*, Urdu *ur*), Afroasiatic (Arabic *ar*), Basque (*eu*), Japonic (Japanese *ja*), Koreanic (Korean *ko*), Niger-Congo (Swahili *sw*), Tai-Kadai (Thai *th*), Turkic (Turkish *tr*), Austroasiatic (Vietnamese *vi*), Sino-Tibetan (Chinese *zh*). Seven of these languages are evaluated in at least three different tasks.

4 Experiments

We evaluate a wide range of multilingual sentence encoders which learn contextual representations. The evaluated models represent a broad spectrum of model sizes, in order to allow practitioners to estimate the trade-off between model size and accuracy.

4.1 Sentence Encoders

The sentence encoders evaluated in the current paper are described in detailed below, and their characteristics summarized in Table 2.

Multilingual BERT (mBERT): mBERT is a transformer-based language model trained with masked language modelling and next sentence prediction objectives similar to the original English BERT model (Devlin et al., 2019)[5]. mBERT is pretrained on the Wikipedias of 104 languages with a shared word piece vocabulary. As discussed in Section 2, its input is not marked with any language-specific signal and mBERT does not have any objective to encode different languages in the same space.

distilmBERT: distilmBERT is a compressed version of mBERT obtained via model distillation (Sanh et al., 2019). Model distillation is a compression technique where a smaller model, called *student*, learns to mimic the behavior of the larger model, called *teacher*, by matching its output distribution. distilmBERT is claimed to reach 92% of mBERT's performance on XNLI while being two times faster and 25% smaller.[6] However, to the best of our knowledge, there is not any comprehensive analysis of distilmBERT's zero-shot performance.

XLM: XLM is a transformer-based language model aimed at extending BERT to cross-lingual setting (Conneau and Lample, 2019). To this end,

---

[5] https://github.com/google-research/bert/blob/master/multilingual.md
[6] https://github.com/huggingface/transformers/tree/master/examples/distillation

Task	Training data	$|train|$	$|test|$	#langs	metric
RST	RST DT	17K	603 – 6,902	6	acc
PDTB	PDTB3	17K	194 – 1,366	7	F_1
X-stance	X-stance-DE	33K	1,446 – 6,153	4	F_1
NLI	MultiNLI	433K	5,010	14	acc
Q/A	Squad 1.1	100K	1,190	11	ex. match/F_1

Table 1: Summary of the datasets used in experiments. "Corpus name-(lang.code)" refers to the part of the corpus belonging to the respective language. #langs refers to the number of *zero-shot* languages, excluding the training language.

XLM increases the shared vocabulary across languages via shared byte pair encoding (BPE) vocabulary. Moreover, unlike BERT, the input sentences are accompanied by language embeddings. There are several different XLM models which differ at either number of training languages or training objectives. In the current study, we consider the following three:

- *XLM-mlm*: The XLM model which is trained with BERT's masked language model (MLM) objective on the Wikipedias of the 15 XNLI languages.

- *XLM-tlm*: In addition to the MLM, this XLM model has a novel training objective which is called Translation Language Model (TLM). In TLM, the model receives a pair of translationally equivalent sentences and tries to predict the masked word by attending both sentences. Hence, the model tries to predict the masked word by looking at its context in another language which encourages representations of different languages to be aligned. TLM is shown to lead a significant increase on XNLI (Conneau and Lample, 2019). XLM-tlm is also trained for 15 XNLI languages but only on parallel data.

- *XLM-100*: This version is trained, like mBERT, on Wikipedia data covering 100 languages using only an MLM objective. Unlike previous XLM models, this version does not utilize language embeddings.

XLM-RoBERTa (XLM-R): XLM-RoBERTa is not an XLM model, in spite of what its name suggests. XLM-R does not use language embeddings, applies sentence-piece tokenization instead of BPE and is not trained on a parallel corpus unlike the XLM-tlm. Instead, it is a RoBERTa model (Liu et al., 2019), which is an optimized version of

BERT, trained on 2.5 TB of cleaned CommonCrawl data covering 100 languages (Conneau et al., 2020). There are two released XLM-R models, XLM-R_{base} and XLM-R_{large}, named after the BERT-architecture they are based on. Compared to original multilingual-BERT, XLM-RoBERTa models have a considerably larger vocabulary size which results in larger models.

4.2 Experimental Setup

A summary of the datasets used in the experiments is provided in Table 1. Except PDTB, all datasets are publicly available. As stated earlier, the training language is English for all tasks except stance detection where German is preferred due the size of the available data. In the spirit of real zero-shot transfer, the validation sets only consist of instances in the training language; hence, no cross-lingual information whatsoever is utilized during training/model selection. For the evaluation metrics, we stick to the default metrics of each task (Table 1).

We set the sequence length to 384 for question answering and RST relation classification; to 250 for stance detection and to 128 for the remaining tasks. At evaluation time, we keep the same configuration. For all models, adam epsilon is set to 1e-8 and maximum gradient norm to 1.0. The learning rate of 2×10^{-5} is used for all the models except XLM-R-large and XLM-100 where it is set to 5×10^{-6}. We adopt the standard fine-tuning approach and fine-tune all models for 4 epochs. We do not apply any early stopping and use the model with the best validation performance during zero-shot experiments. All tasks are implemented using Huggingface's Transformers library (Wolf et al., 2019). As fine-tuning procedure is known to show high variance on small training datasets, all models are run for 4 times with different seeds and the average performance is reported. For XLM and XLM-tlm models, we fall back to English lan-

Model	Langs	Parameter count	Vocab. size	# of layers
distilmBERT	104	134M	30K	6
mBERT	104	177M	30K	12
XLM-mlm	15	250M	95K	12
XLM-tlm	15	250M	95K	12
XLM-100	100	570M	200K	16
XLM-R$_{base}$	100	270M	250K	12
XLM-R$_{large}$	100	550M	250K	24

Table 2: The characteristics of the sentence encoders evaluated in the experiments
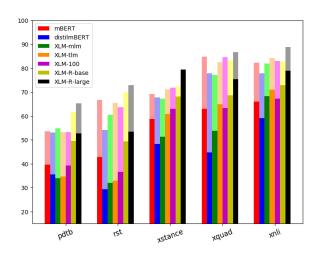

Figure 1: Overview of performance of each sentence encoder on all Disco-X tasks. The semi-transparent bars represent source language performance (German for X-stance, English for the rest) while the solid bars represent the zero-shot performance, i.e. the mean performance across all languages *except* the training language. All values are averages over independent training runs.

guage embeddings for non-XNLI languages. All experiments are run on a single TITAN X (12 GB) GPU.

5 Results and Discussion

We provide an overview of the main results in Figure 1. The detailed results with per-language breakdown are provided in the Appendix A.

Overall, there is a clear difference between the training and zero-shot performance of all models. When averaged over all tasks, the performance loss in zero-shot transfer ranges from 15.58% (XLM-R-large) to 34.96% (distilmBERT) which clearly highlights the room for improvement, especially with smaller model sizes. In the rest of the section, we discuss the results in terms of the encoder type, task and the languages.

Model-wise analysis The ranking of the encoders displays relatively little variation across tasks, with XLM-R$_{large}$ exhibiting the best zero-shot performance across all tasks by outperforming the second best model (XLM-R$_{base}$) by 5.98%. distillmBERT, on the other hand, fails to match the performance of other encoders.[7]

The Translation Language Model (TLM) objective is proved to be a better training objective than MLM by consistently outperforming the vanilla XLM in all tasks. XLM-tlm outperforms XLM-100 on XNLI languages as well which is possibly because of the 'curse of multilinguality' (Conneau et al., 2020), the degradation of the overall performance in proportion to the number of languages in the training. However, training setting (e.g. training data, hyperparameters) outplays the 'curse of multilinguality' as XLM-R$_{base}$ clearly outperforms XLM-tlm even on XNLI languages. It would be interesting to see how an XLM-R trained with TLM objective on small set of languages, e.g. XNLI languages, would perform.

DistillmBERT is the lightest model evaluated in the current investigation. It is shown to retain 92% of the mBERT's performance on certain XNLI languages.[8] The results suggest that distillmBERT delivers its promise, although to a lesser extent. When averaged over all tasks, distillmBERT retains 93% of the source language performance of mBERT. However, its relative performance significantly drops to 82% on zero-shot transfer. That is, distillmBERT is not as successful when it comes to copying mBERT's cross-lingual abilities. Furthermore, its performance (relative to mBERT) is not stable across tasks either. It only achieves 69% of

---

[7]The only exception is the XLM and XLM-tlm's performance on non-XNLI languages where distillmBERT manages to outperform them but not always by a large margin.

[8]https://github.com/huggingface/transformers/tree/master/examples/distillation

mBERT's zero-shot performance on RST whereas 89% on XNLI. The low memory requirement and its speed (with the same batch size, it is x2 faster than mBERT and x5 than XLM-R$_{large}$) definitely makes distillmBERT a favorable option; however, the results show that its zero-shot performance is considerably lower than its source language performance and is highly task-dependent, hence, hard to predict.

Task-wise Analysis Table 3 shows to what extent encoders manage to transfer their source language performance to zero-shot languages. Overall, the zero-shot performances show high variance across tasks which is quite interesting given that all tasks are on the same linguistic level. It is also surprising that mBERT manages a better zero-shot transfer performance than all XLM models while being almost as consistent as XLM-R$_{base}$.

Overall, the results show that even modern sentence encoders struggle to capture inter-sentential interactions in both monolingual and multilingual settings, contrary to the what the high performances on well-known datasets (e.g. PAWS (Hu et al., 2020)) may suggest. We believe that this finding supports our motivation to propose new probing tasks to have a fuller picture of the capabilities of these encoders.

Language-wise Analysis: In all tasks, regardless of the model, training-language performance is better than even the best zero-shot performance. The only exception is the XLM-R-large's performance on the X-stance where the zero-shot performance is on par with its performance on the German test set.

An important aspect of cross-lingual research is predictability. The zero-shot performance of a certain language do not seem to be stable across tasks (e.g. German is the language with the worst RST performance; yet it is one of the best in XNLI). We further investigate this following Lauscher et al. (2020), who report high correlation between syntactic similarity and zero-shot performance for low-level tasks, POS-tagging and dependency parsing. We conduct the same correlation analysis using Lang2Vec (Littell et al., 2017). However, syntactic and geographical similarity only weakly correlates with zero-shot performances across the tasks (Pearson's $r = .46$ and Spearman's $r = .53$ on average for syntactic; Pearson's $r = .30$ and Spearman's $r = .45$ for geographical similarity). Such low correlations are important as it further supports the claim that the tasks are beyond the sentence level and also highlights a need for further research to reveal the factors at play during zero-shot transfer of discourse-level tasks.

6 Conclusion

As pre-trained multilingual sentence encoders have become prevalent in natural language processing, research on cross-lingual zero-shot transfer gains increasing importance (Hu et al., 2020; Liang et al., 2020). In this work, we evaluate a wide range of sentence encoders on a variety of discourse-level tasks in a zero-shot transfer setting. Firstly, we enrich the set of available probing tasks by introducing three resources which have not been utilized in this context before. We systematically evaluate a broad range of widely used sentence encoders with considerably varying sizes, an analysis which has not been made before.

The main variable we look at is the performance gap between training-language evaluation and zero-shot evaluation. Unsurprisingly, nearly always there is such a gap, but its magnitude depends on a number of factors:

- **Distillation**: the distilled mBERT model has a larger gap than the full mBERT model, indicating loss of multilingual transfer ability during distillation.

- **Language similarity**: the gap correlates only weakly with measures of language similarity (syntactic and geographical), indicating that sentence encoders generally transfer discourse-level information about as well between similar and dissimilar languages.

- **High variance**: apart from the above, we also observe a generally high variance in the gap magnitude between different tasks in our benchmark suite.

These observation provide several starting points for future work: investigating why knowledge distillation seems to hurt zero-shot performance to a much greater extent than same-language sentence encoding ability and what can be done to solve this problem, and explaining the large variations in the zero-shot transfer gap between different discourse-level NLP tasks.

Model	PDTB	RST	X-stance	XQuAD	MNLI	Average $\pm$ std
mBERT	74.49	64.18	84.75	74.22	80.28	75.58 ± 6.92
distilmBERT	66.13	54.37	71.34	57.35	75.9	65.02 ± 8.15
XLM-mlm	60.32	52.93	76.4	69.68	83.47	68.56 ± 10.93
XLM-tlm	63.49	50.36	85.57	78.76	84.26	72.49 ± 13.56
XLM-100	73.76	57.54	87.62	74.89	81.01	74.96 ± 10.02
XLM-R$_{base}$	78.96	70.75	94.29	82.44	88.1	82.91 ± 8.00
XLM-R$_{large}$	79.91	73.33	100.4	86.81	89	85.89 ± 9.11

Table 3: Relative zero-shot performance of each encoder to the source language performance (metrics differ between tasks but higher is better in all cases). The figures shows what percentage of the source language performance is retained through zero-shot transfer in each task. Hu et al. (2020) refer to this as the *cross-lingual transfer gap*. A score above 100 indicates that a better zero-shot performance than that of training.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The rst spanish-chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166.

Paula CF Cardoso, Erick G Maziero, Maria LC Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia HM Rino, Maria das Gracas Volpe Nunes, and Thiago AS Pardo. 2011. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019a. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019b. Seeing things from a different angle: Discovering diverse perspectives about claims. 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 542–557.

Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informaç oes discursivas visandoa sumarizaç ao automática. *Proceedings of TIL*.

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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Iria Da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the rst spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In International Conference on Machine Learning, pages 4411–4421. PMLR.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494.

Mikel Iruskieta, Mara Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The rst basque treebank: an online search interface to check rhetorical relations. In 4th Workshop" RST and Discourse Studies", Brasil, October, pages 21–23.

Murathan Kurfalı and Robert Östling. 2019. Zero-shot transfer for implicit discourse relation classification. In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pages 226–231.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7315–7330, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6008–6018.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber, and Deyi Xiong. 2020. Shallow discourse annotation for chinese ted talks. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 1025–1032.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text & Talk, 8(3):243 – 281.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4497–4510.

Giedre Valunaite Oleskeviciene, Deniz Zeyrek, Viktorija Mazeikiene, and Murathan Kurfalı. 2018. Observations on the annotation of discourse relational devices in ted talk transcripts in lithuanian. In Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI, volume 2155, pages 53–58.

Thiago Alexandre Salgueiro Pardo and Eloize Rossi Marques Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. Anais do V Encontro de Corpora, pages 24–25.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001.

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Towards building a discourse-annotated corpus of russian. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies" Dialogue.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, pages 1–26.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, and Hugging Face. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Manfred Stede. 2004. The potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102.

Sara Tonelli and Elena Cabrio. 2012. Hunting for entailing pairs in the penn discourse treebank. In *Proceedings of COLING 2012*, pages 2653–2668.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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 833–844.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19.

Jianhua Yuan, Yanyan Zhao, Jingfang Xu, and Bing Qin. 2019. Exploring answer stance detection with recurrent conditional attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7426–7433.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (tedmdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.

A   Task-wise Results

Model	en	de	es	eu	pt	ru	zh	AVG
mBERT	66.7	29.2	39.3	31.1	58.6	48.0	50.7	42.8
distilmBERT	54.1	16.3	25.7	21.4	44.5	32.2	36.5	29.4
XLM-mlm	60.6	25.5	33.2	14.1*	40.4*	39.8	39.4	32.1
XLM-tlm	65.5	26.0	35.2	13.3*	42.0*	39.9	41.3	33.0
XLM-100	63.8	24.3	34.6	26.2	55.2	40.0	39.8	36.7
XLMR-b	69.8	37.6	44.7	39.4	61.9	56.2	56.7	49.4
XLMR-l	72.9	44.8	46.8	47.0	65.6	59.3	57.3	53.5

Table 4: RST zero-shot results (Accuracy) for each language. * denotes that the language is not one of the training languages of the respective sentence encoder.

Model	en	de	lt	pl	pt	ru	tr	zh	AVG
mBERT	53.6	42.7	39.2	33.9	46.7	33.1	40.3	43.5	39.9
distilmBERT	53.1	42.7	30.0	34.7	41.1	32.6	29.4	35.4	35.1
XLM-mlm	54.9	44.9	19.5*	20.6*	28.9*	33.8	43.5	40.5	33.1
XLM-tlm	53.3	45.9	20.1*	21.3*	26.8*	37.1	41.9	43.6	33.8
XLM-100	54.6	41.9	41.6	32.5	44.5	34.2	35.9	40.4	38.7
XLMR-b	61.8	49.5	49.6	40.4	53.5	42.7	54.4	51.4	48.8
XLMR-l	65.4	53.4	49.4	42.8	59.5	48.9	53.8	58.1	52.3

Table 5: PDTB zero-shot results (F_1) for each language. * denotes that the language is not one of the training languages of the respective sentence encoder.

Model	de	en	fr	it	zh	AVG
mBERT	69.3	60.2	60.7	63.2	50.8	58.7
distilmBERT	67.7	49.8	48.7	59.5	35.2	48.3
XLM-mlm	67.3	52.6	55.0	56.2*	41.8	51.4
XLM-tlm	71.2	60.4	62.5	59.6*	61.1	60.9
XLM-100	71.8	62.3	64.8	64.0	60.6	62.9
XLMR-b	72.3	65.8	70.4	69.9	66.7	68.2
XLMR-l	79.3	80.9	79.0	78.9	79.5	79.6

Table 6: X-stance zero-shot results (F_1) for each language. * denotes that the language is not one of the training languages of the respective sentence encoder.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	AVG
mBERT	82.3	65.7	69.4	72.1	68.2	75.9	75.3	60.6	69.8	51.3	54.7	62.2	58.8	70.9	69.7	66.1
distilmBERT	77.9	60.3	63.9	65.7	61.4	70.1	69.9	54.7	63.6	46.6	39.1	57.3	54.1	59.2	62.4	59.2
XLM-mlm	81.9	68.5	73.7	73.0	73.3	75.3	75.2	64.4	72.0	64.9	49.2	67.3	62.8	70.3	67.3	68.4
XLM-tlm	84.2	71.1	76.5	76.2	74.3	78.3	77.9	66.5	75.3	67.4	53.9	70.8	62.7	72.8	69.3	70.9
XLM-100	83.1	67.9	72.6	73.3	72.4	76.6	75.5	64.7	71.3	58.4	39.7	68.2	62.0	72.7	67.0	67.3
XLMR-b	82.8	71.0	77.3	75.7	75.3	78.2	76.9	68.6	75.2	66.4	71.6	72.4	65.2	74.6	73.0	73.0
XLMR-l	88.8	78.6	83.0	82.9	81.8	84.5	82.7	76.0	79.3	71.6	77.0	78.7	71.5	79.5	79.3	79.0

Table 7: XNLI zero-shot results (Accuracy) for each language

Model	en	ar	de	el	es	hi
mBERT	84.8/72.9	62.6/46.0	72.5/56.8	64.4/47.1	75.3/56.3	58.6/45.1
distilmBERT	78.0/65.9	44.6/28.3	57.6/41.0	37.6/21.2	60.5/40.0	34.9/20.5
XLM-mlm	77.2/64.5	59.9/43.2	66.0/50.4	57.8/39.5	67.7/49.8	47.5/33.0
XLM-tlm	82.5/70.4	68.1/51.6	73.7/57.6	69.5/51.2	77.1/59.2	65.6/50.2
XLM-100	84.6/73.4	67.6/50.3	73.6/58.3	63.9/45.1	77.3/59.1	60.2/44.5
XLMR-b	83.3/72.4	65.0/47.1	73.4/57.6	71.9/54.5	75.5/57.1	68.3/50.9
XLMR-l	86.8/75.5	74.1/55.6	79.5/62.6	79.8/61.4	82.0/62.3	75.4/58.6
Model	ru	th	tr	vi	zh	AVG
mBERT	71.4/54.9	43.3/34.4	54.8/40.8	68.1/48.9	58.3/48.2	62.9/47.8
distilmBERT	58.9/40.2	20.9/13.9	37.9/21.8	47.5/28.2	46.9/33.8	44.7/28.9
XLM	64.1/47.0	24.9/12.4	50.2/34.6	60.3/41.3	39.8/30.1	53.8/38.1
XLM-tlm	72.6/55.3	33.3/21.9	65.0/47.5	71.8/51.3	53.4/43.8	65.0/48.9
XLM-100	73.7/57.6	22.4/13.6	66.7/49.9	73.9/54.8	54.1/44.5	63.3/47.8
XLMR-b	73.3/56.9	67.1/55.5	67.5/50.4	73.0/53.4	51.6/41.7	68.7/52.5
XLMR-l	79.4/62.9	73.7/62.6	74.7/58.5	79.4/59.4	55.5/46.7	75.4/59.1

Table 8: XQuAD results (F_1/Exact-match) for each language