

What the [MASK]? Making Sense of Language-Specific BERT Models

Debora Nozza, Federico Bianchi and Dirk Hovy

Bocconi University

Via Sarfatti 25, 20136 Milan

{debora.nozza, f.bianchi, dirk.hovy}@unibocconi.it

Abstract

Recently, Natural Language Processing (NLP) has witnessed an impressive progress in many areas, due to the advent of novel, pretrained contextual representation models. In particular, Devlin et al. (2019) proposed a model, called BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers), which enables researchers to obtain state-of-the-art performance on numerous NLP tasks by fine-tuning the representations on their data set and task, without the need for developing and training highly-specific architectures. The authors also released multilingual BERT (mBERT), a model trained on a corpus of 104 languages, which can serve as a universal language model. This model obtained impressive results on a zero-shot cross-lingual natural inference task. Driven by the potential of BERT models, the NLP community has started to investigate and generate an abundant number of BERT models that are trained on a particular language, and tested on a specific data domain and task. This allows us to evaluate the true potential of mBERT as a universal language model, by comparing it to the performance of these more specific models. This paper presents the current state of the art in language-specific BERT models, providing an overall picture with respect to different dimensions (i.e. architectures, data domains, and tasks). Our aim is to provide an immediate and straightforward overview of the commonalities and differences between Language-Specific (language-specific) BERT models and mBERT. We also provide an interactive and constantly updated website that can be used to explore the information we have collected, at <https://bertlang.unibocconi.it>.

1 Introduction

In all natural languages, word meaning varies with and is determined by context, and one of the main challenges of Natural Language Processing (NLP) has been (and remains) to model this property of meaning. Embedding-based language models (Mikolov et al., 2013) have been shown to capture word meaning more efficiently than previous methods, allowing for both qualitative analysis of similarities and improved performance when used as input to predictive models. However, while embeddings represent word *types* based on their general contextual co-occurrences, they do not learn context-specific representations for each word *token*.

Recently, NLP has witnessed the advent of a groundbreaking new language model developed by Google researchers, called Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). It learns contextual representations for word tokens, thereby getting at their contextual variation in meaning. Contextualized BERT embeddings have since also dominated the leaderboards in a wide variety of NLP tasks.

The power of BERT representations lies in the fact that it is essentially a pretrained model that can be fine-tuned over specific downstream tasks, which enables it to achieve state-of-the-art results. The fundamental underlying component of this architecture is the Transformer model (Vaswani et al., 2017), an attention-based mechanism that has been shown to be effective in many different tasks. Both the Transformer and BERT have gathered much attention, and there is now a wealth of research articles and blog posts describing the inner workings of these models (Rogers et al., 2020, among others).

Given the overwhelming success of BERT, a multilingual BERT model (mBERT)¹ has been proposed, supporting over 100 languages, including Arabic, Dutch, French, German, Italian, or Portuguese. The model is trained on different domains, like social media posts or newspaper articles. mBERT has shown great capabilities in zero-shot cross-lingual tasks (Pires et al., 2019).

Due to the remarkable results of these models, an abundant number of BERT model extension has recently been introduced by researchers and industry practitioners from several countries: Currently, there are around 5k repositories mentioning “bert” on GitHub.com, and we can expect further demand for BERT extensions. These models are trained on a particular language and tested on a specific data domain and task, with the promise of maximizing performance across more tasks in that language, saving other users further fine-tuning.

However, it has so far not been clearly demonstrated whether the advantage of training a language-specific model is worth the expense in terms of computational resources², rather than using the *unspecific* multilingual model.

Moreover, the NLP community is now facing a problem organizing the plethora of models that are being released. These models are not only trained on different data sets, but also use different configurations and architectural variants. To give a concrete example, the original BERT model was trained using the WordPiece tokenizer (Wu et al., 2016), however, a recent language-specific model (CamemBERT (Martin et al., 2019)) used the SentencePiece tokenizer (available as OSS software) (Kudo and Richardson, 2018).

Identifying which model is the best for a specific task, and whether the mBERT model is better than language-specific models is a key step future progress in NLP, and will impact the use of computational resources. Surveying both GitHub and the literature, we identified 30 different pretrained language-specific BERT models, covering 18 Languages and tested on 29 tasks, resulting in 177 different performance results (Le et al., 2019; Antoun et al., 2020; Martin et al., 2019; Alabi et al., 2019; Kuratov and Arkhipov, 2019; Arkhipov et al., 2019; Virtanen et al., 2019; Polignano et al., 2019; de Vries et al., 2019; Cui et al., 2019). We outline some of the parameters here, and introduce the associated website for up-to-date searches. We hope to give NLP researchers and practitioners a clear overview of the tradeoffs before approaching any NLP task with such a model.

The contributions of this paper are the following:

1. we present an overall picture of language-specific BERT models from an architectural, task- and domain-related point of view;
2. we summarize the performance of language-specific BERT models and compare with the performance of the multilingual BERT model (if available);
3. we introduce a website to interactively explore state-of-the-art models. We hope this can serve as a shared repository for researchers to decide which model best suits their needs.

2 Bidirectional Encoder Representations from Transformers

We assume that most readers who are interested in the topic have a basic understanding of BERT and its components. However, for completeness’ sake, we include a brief and high-level overview of the most important aspects here.

2.1 BERT

BERT uses the Transformer (Vaswani et al., 2017) architecture to learn word embeddings. The Transformer is a recent architectural advancement that can be included in deep networks for sequence modeling. Instead of modeling sequences as RNNs or LSTMs, the Transformer learns global dependencies between input and output, using only attention mechanisms.

¹<https://github.com/google-research/bert/blob/master/multilingual.md>

²These models require a large amount of computational resources unaffordable for many users, and comes with severe ecological costs: training BERT on a GPU is roughly equivalent to a trans-American flight in terms of CO2 emissions (Strubell et al., 2019)).

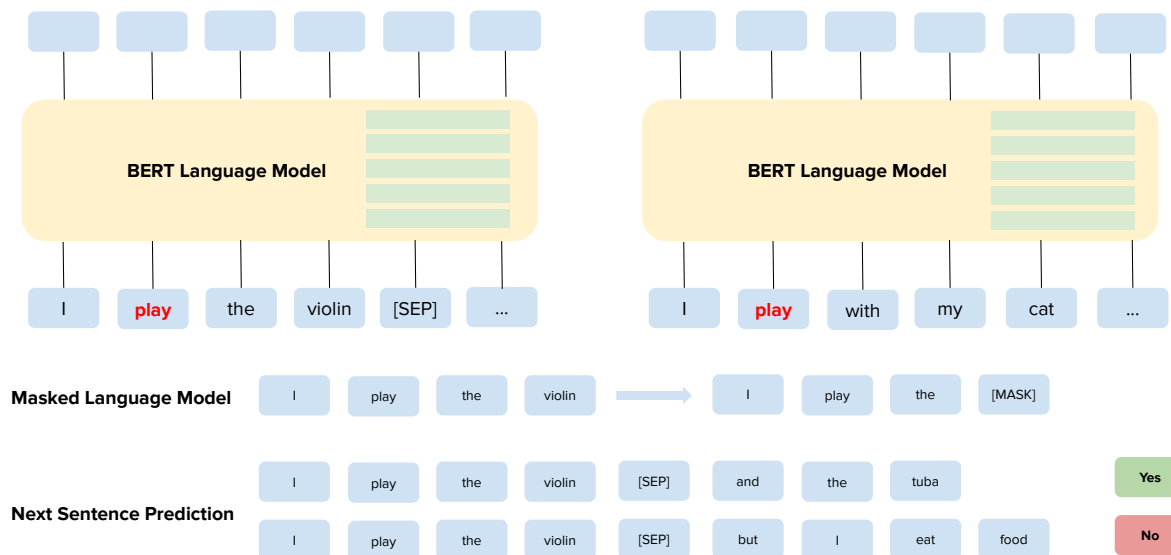


Figure 1: A schematic representation of BERT, masked language model and next sentence prediction. Different words have different meanings and BERT looks at the word context to generate contextual representations.

Transformers greatly shifted the focus of the research community towards attention-based architectures. The encoder-decoder structure based on transformers is also incorporated into BERT.

Devlin et al. (2019) introduced BERT in 2018 as a context-sensitive alternative to previous word embeddings (which assume a word always has the same representation, independent of its context). The model essentially stacks several encoder-decoder structures based on transformers together. It uses masks to blank out individual words, forcing the model to “fill in the blanks”, thereby increasing its context-sensitivity. Two key elements in the BERT pretraining process are the masked language model and the next sentence prediction. In the former process, a random subsample (in the BERT paper, 15%) of the words in a text are replaced by a [MASK] token, and the task is to predict the correct token. The latter process instead is the task of predicting how likely one sentence is to follow another one in text. See Figure 1 for a schematic view on BERT. Other than traditional word embeddings, BERT representations are not a fixed lookup table, but require the full context to produce a word representation. The vocabulary is defined in advance and it is based on WordPiece (Wu et al., 2016), a tokenization algorithm that generates sub-word tokens.

Due to its size in terms of parameters, the model usually comes in a pretrained format, which can be fine-tuned on the task or data set. Simple classification layers can be stacked on top of the pretrained BERT to provide predictions for several tasks such as sentiment analysis or text classification.

2.2 Multilingual BERT, ALBERT and RoBERTA

Subsequently, BERT was extended to include several languages. Multilingual (mBERT) was part of the original paper (Devlin et al., 2019), and is pretrained over several languages using Wikipedia data. This allows for zero-shot learning across languages, i.e., training on data from one language and applying the model to data in another language.

Along the same lines, Lan et al. (2019) introduced A Lite BERT (ALBERT), to reduce the computational needs of BERT. ALBERT includes two parameters reduction techniques, that reduce the number of trainable parameters without significantly compromising performance. Moreover, the authors introduce another self-supervised loss, related to sentence order prediction that is meant to address the limits of next sentence prediction used in the BERT model. Another recent paper (Liu et al., 2019) has shown that BERT is sensitive to different training strategies; the authors introduce RoBERTA (Liu et al., 2019) as a well-optimized version of BERT.

Language	Model	NLP Task	Dataset	Dataset-Domain	Measure	Performance	mBERT	Difference with mBERT	Source
Dutch	BERTje	NER	CoNLL-2002	news	F1 (test)	88.3	80.7	7.6	
English	BERT-Base	PI	MRPC	news	F1 (dev)	88.9	N/A	N/A	
English	BERT-Base	QA	SQuAD v1.1	wiki	F1 (dev)	88.5	N/A	N/A	
English	ALBERT (1M)	TER	RTE	news, wiki	Accuracy (dev)	88.1	N/A	N/A	
Finnish	FinBERT	TC	Yle news	news	Accuracy (test size 10K)	90.57	88.44	2.13	
French	CamemBERT	NER	French Treebank	news	F1 (test)	88.39	87.52	0.87	
German	deepset-GermanBERT	TC	10kGNAD	news	Accuracy	90.5	88.8	1.7	
Italian	Italian BERT (XXL)	NER	I-CAB 2009	news	F1	88.13	85.18	2.95	
Spanish	BETO	NER	CoNLL 2000, 2002, 2007	news	F1	88.43	87.38	1.05	

Figure 2: The BertLang website front-end interface.

3 Making-Sense of Language-Specific BERT Models

While multi- and cross-lingual BERT representations allow for zero-shot learning and capture universal semantic structures, they do gloss over language-specific differences. Consequently, a number of language-specific BERT models have been developed to fill that need. These models almost always showed better performance on the language they were trained for than the universal model.

In order to navigate this wealth of constantly changing information, a simple overview paper is no longer sufficient. While we aim to give a general overview here, we refer the interested reader to the constantly updated online resource, BertLang.

3.1 BertLang

We introduce BertLang (<https://bertlang.unibocconi.it>), a website where we have gathered different language-specific models that have been introduced on a variety of tasks and data sets. Most of the models are available as GitHub links, and some of them are described in research papers, but very few have been published in peer-reviewed conferences³. In addition to providing a searchable interface, BertLang also provides the possibility to add new information. While we hope to independently verify the reported results in the future, for now, we only list the various models and conditions.

We open-source both data and code to build the website⁴, this will make it possible for other researchers to contribute to the collection of language-specific BERT models.

Figure 2 shows the frontend page of our website, showing a table that contains languages, tasks, and performances of different models. We also provide links to the references and code from which we retrieved that information. Beyond this information, we report the performance evaluation metric, the average performance obtained by the language-specific, and – where available – the corresponding performance of mBERT model and their difference.

³We do not include resources that feature only a model without reporting any performance results.

⁴<https://github.com/MilaNLP/bertlang>

Task	Metric	Avg. lang-specific BERT	Avg. mBERT	Diff.
Named Entity Recognition	F1	85.26	80.87	4.39
Natural Language Inference	Accuracy	78.35	74.60	3.75
Paraphrase Identification	Accuracy	88.44	87.74	0.70
Part of Speech Tagging	Accuracy	97.06	95.87	1.19
Part of Speech Tagging	UPOS	98.28	97.33	0.95
Sentiment Analysis	Accuracy	90.17	83.80	6.37
Text Classification	Accuracy	88.96	85.22	3.75

Table 1: Summary of average performance of different language-specific BERT models on various tasks.

3.2 Language-Specific BERT Models

The models we index vary along with a number of dimensions, which we discuss below. The main distinction, however, is the specific language the model was trained on. The availability of data sets in that language determines the tasks and domains this model was applied to.

Table 1 shows a summary of the results for the most frequent NLP tasks investigated across several languages. The results clearly show that on average, language-specific BERT models obtain higher results with respect to mBERT in all the considered tasks. However, while this holds for averages, with the proliferation of languages, tasks, and data sets, there is a huge variation in the individual performances. In the following, we analyze the possible views of the collected results in more detail.

Languages Covered The language-specific BERT models proposed range from languages that have a high number of resources available on the web for training (e.g., French, Italian) to low-resource languages, such as Yorba and Mongolian. At the current date, we are covering 18 languages.

Interestingly, from the results it is possible to grasp that low-resources languages (e.g., Yorba and Arabic) are actually the ones with the highest improvement with respect to mBERT. Since mBERT is trained on Wikipedia, this finding can probably be explained by the fact that developers of language-specific BERT models are more likely to be experts on other resources for that language, or to collect more data. This makes a greater difference for low-resource languages.

Architectures The most popular architecture is the standard BERT one, but lately, the introduction (and the good performances) of both ALBERT and RoBERTa has made researchers consider those two latter models as well to pretrain language models.

RoBERTa has been used as the base model for the French CamemBERT (Martin et al., 2019), as well as the Italian Gilberto⁵ and Umberto⁶.

mBERT was used to initialize and fine-tune models for languages such as Russian (Kuratov and Arkhipov, 2019), Slavic languages (Arkhipov et al., 2019)⁷ and Yorba (Alabi et al., 2019). The latter is a noteworthy example of how the scarcity of available data in low resource languages can be overcome. Fine-tuning mBERT instead of pretraining from scratch allowed the authors to produce a model without access to large amounts of data.

NLP tasks We currently index results for 29 NLP tasks. Table 1 reports the results for the most popular tasks in the collected data, with Named Entity Recognition (NER) the most frequent task (22 entries). Looking at the source of the test data (see the released website for the complete information), we observe that there are some multilingual benchmark data sets that are used for the same NLP task in different languages. Some of them have been released by research group publishing in well-known NLP conferences (Yang et al., 2019; Sanguinetti and Bosco, 2015; Conneau et al., 2018; Völker et al., 2019), while others have been released in conjunction with shared tasks such as SemEval or CoNLL (Zeman et al., 2018;

⁵<https://github.com/idb-ita/GilBERTo>

⁶<https://github.com/musixmatchresearch/umberto>

⁷Here “Slavic” includes Russian, Bulgarian, Czech and Polish.

Navigli et al., 2013; Bosco et al., 2016; Benikova et al., 2014). The latter group shows the effect shared task have on providing the NLP community with benchmark references.

Remarkably, the noun sense disambiguation task is the only task where language-specific BERT performances are lower than the mBERT ones. As stated by the authors (Le et al., 2019), this could be due to the fact that the training corpora have been machine-translated from English to French, making mBERT probably better suited for the task than a model trained on native French.

Sentiment analysis is the task where language-specific BERT models obtain the highest improvements with respect to mBERT. Following the previous intuition, for Arabic (Antoun et al., 2020) this can be explained by considering the peculiar language of the test data set, which demonstrates the ability of the language-specific AraBERT model to handle dialects — even if they were not explicitly included in the training set.

Beyond the well-known NLP tasks, it is interesting to note that language-specific tasks have been investigated as well, e.g., the Die/Dat (gendered determiners) disambiguation task in Dutch (Delobelle et al., 2020), obtaining impressive improvements with respect to state-of-the-art (Allein et al., 2020) (~ 23% points accuracy improvement).

Domains There is a huge variety of domains considered in language-specific BERT models. We need to make a distinction, though, between data sets used to pretrain the models and data sets used to evaluate the models.

Data used for training mainly varies across three source corpora: (i) Wikipedia, (ii) OPUS Corpora (Tiedemann, 2012) and (iii) OSCAR (Ortiz Suárez et al., 2019). Wikipedia is currently comprising more than 40 million articles created and maintained as an open collaboration project in 301 different languages, making it the largest and most popular multilingual online encyclopedia. mBERT, for example, was trained over 100 different language-specific Wikipedia versions. OPUS is a freely available collection of parallel corpora, covering over 90 languages. The largest domains covered by OPUS are legislative and administrative texts, translated movie subtitles and localization data from open-source software projects (Tiedemann, 2012). OSCAR (Open Super-large Crawled Almanach coRpus) (Ortiz Suárez et al., 2019) is a huge multilingual corpus obtained by filtering the Common Crawl corpus, which is a parallel multilingual corpus comprised of crawled documents from the internet.

Several models concatenate more sources to have enough data to pretrain BERT, for example BERTje (Dutch BERT), which concatenates news, book data, and Wikipedia data and other text. Languages with more limited availability of data, such as Yorub, have brought researchers to fine-tune mBERT instead of pretraining from scratch. A notable case is the Italian BERT model ALBERTO (Polignano et al., 2019), which is the only one that has been trained only on social media data (specifically, on 2 million Twitter posts in Italian language).

On the other hand, different domain data sets have been used to evaluate the models; these range from review data for sentiment analysis tasks to transcripts and news for more traditional tasks, such as part of speech tagging. News data are the most common domain, presumably because they are easier to retrieve, and because their more formal register makes them more suited for tasks such as part of speech tagging, dependency parsing, and named entity recognition. Similarly, social media posts from Twitter are mostly used in tasks like sentiment analysis and identification of offensive language.

4 Conclusions

BERT (Devlin et al., 2019) has greatly improved results in many different NLP tasks and has become a mainstay of the community. Following this development, a multilingual BERT and several language-specific versions have been developed and contributed even more to the success of NLP applications.

In this paper, we have analyzed the current state-of-the-art, showing languages are covered, which tasks tackled, and which domains considered in pretrained language-specific BERT models. Moreover, we have underlined the huge variability models and the difficulty for researchers to find the best model for a specific task, language, and domain. To this end, we have introduced BertLang, a website that allows researchers to search and explore the current state-of-the-art with respect to language-specific BERT models.

In the future, we plan to provide independent verification of reported results and direct comparisons of language-specific BERT models on specific domains and tasks. We plan to use the same data to fine-tune the models providing comparable performance values for the models. We believe these comparisons will be beneficial to the community of both researchers and beginning practitioners in NLP.

References

- Jesujoba O Alabi, Kwabena Amponsah-Kaakyire, David I Adelani, and Cristina España-Bonet. 2019. Massive vs. curated word embeddings for low-resourced languages. The case of Yorùbá and Twi. *arXiv preprint arXiv:1912.02481*.
- Liesbeth Allein, Artuur Leeuwenberg, and Marie-Francine Moens. 2020. Binary and multitask classification model for Dutch anaphora resolution: Die/Dat prediction. *arXiv preprint arXiv:2001.02943*.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93. Association for Computational Linguistics.
- Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. GermEval 2014 named entity recognition shared task: companion paper. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 104–112.
- Cristina Bosco, Tamburini Fabio, Bolioli Andrea, and Alessandro Mazzei. 2016. Overview of the evalita 2016 part of speech on twitter for italian task. In *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2016*, volume 1749, pages 1–7. CEUR Workshop Proceedings (CEUR-WS.org).
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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, EMNLP 2018*, pages 2475–2485. Association for Computational Linguistics.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for Chinese BERT. *arXiv preprint arXiv:1906.08101*.
- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.
- Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a dutch RoBERTa-based language model. *arXiv preprint arXiv:2001.06286*.
- 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, NAACL-HLT 2019*, pages 4171–4186.
- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations*, pages 66–71. Association for Computational Linguistics.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint arXiv:1905.07213*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. FlauBERT: Unsupervised language model pre-training for French. *arXiv preprint arXiv:1912.05372*.

- 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*.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.
- Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval 2013*, pages 222–231. The Association for Computer Linguistics.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora, CMLC-7*. Leibniz-Institut für Deutsche Sprache.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 4996–5001. Association for Computational Linguistics.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. ALBERTO: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the 6th Italian Conference on Computational Linguistics, CLiC-it 2019*. CEUR Workshop Proceedings (CEUR-WS.org).
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.
- Manuela Sanguinetti and Cristina Bosco. 2015. PartTUT: The Turin university parallel treebank. In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, volume 589 of *Studies in Computational Intelligence*, pages 51–69. Springer.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 3645–3650. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218. European Language Resources Association (ELRA).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.
- Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large universal dependencies treebank for German. In *Proceedings of the 3rd Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57. Association for Computational Linguistics.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 2019*, pages 3685–3690. Association for Computational Linguistics.
- Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21. Association for Computational Linguistics.