mALBERT: Is a Compact Multilingual BERT Model Still Worth It?

Christophe Servan[1,2], **Sahar Ghannay**[1], **Sophie Rosset**[1]
[1]Université Paris-Saclay, CNRS, LISN, [2]QWANT
{firstname.lastname}@lisn.upsaclay.fr

Abstract

Within the current trend of Pretained Language Models (PLM), emerge more and more criticisms about the ethical and ecological impact of such models. In this article, considering these critical remarks, we propose to focus on smaller models, such as compact models like ALBERT, which are more ecologically virtuous than these PLM. However, PLMs enable huge breakthroughs in Natural Language Processing tasks, such as Spoken and Natural Language Understanding, classification, Question–Answering tasks. PLMs also have the advantage of being multilingual, and, as far as we know, a multilingual version of compact ALBERT models does not exist. Considering these facts, we propose the free release of the first version of a multilingual compact ALBERT model, pre-trained using Wikipedia data, which complies with the ethical aspect of such a language model. We also evaluate the model against classical multilingual PLMs in classical NLP tasks. Finally, this paper proposes a rare study on the subword tokenization impact on language performances.

Keywords: Transformer, Multilingual, Compact Model, Tokenization

1. Introduction

Recent advances in the field of Natural Language Processing (NLP) are due to the development of transfer learning and the availability of Pre-trained Language Models (PLM) based on Transformer architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019). As they provide contextualized semantic representation, they contribute both to advancing the state-of-the-art on several NLP tasks and also to evolving training practices through the use of fine-tuning.

The recent trend consists of training large PLMs on ever larger corpora with an ever-increasing amount of parameters, which requires considerable computational resources that only a few companies and institutions can afford, such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023) or BLOOM (Scao et al., 2022). This trend raises questions about the temporal, financial, and environmental aspects of these models (Strubell et al., 2019; Moosavi et al., 2020). Therefore, one of the favored tracks is the reduction of computational resources involved while pre-training, fine-tuning, and inference of these models.

As far as we know, compact models, such as the ALBERT model(Lan et al., 2019), are a possible answer since they have been evaluated on the comprehension tasks covered by GLUE (Wang et al., 2018) and the question-answering task with the SQuAD corpus (Rajpurkar et al., 2016) with abundant data. They also have shown their effectiveness on lower-scale learning problems in poorly endowed languages but only in a monolingual context (Lan et al., 2019; Cattan et al., 2021). As far as we know, the multilingual version of such a model does not exist.

All Pre-trained Language Model uses subword unit tokenization in order to alleviate the open vocabulary problem. We take the opportunity of a new language model to conduct a short study of the impact of the subword unit vocabulary. Subword units come from studies conducted in machine translation using compression methods in order to reduce the vocabulary amount and to handle the Out-Vocabulary-Words (Chitnis and DeNero, 2015; Schuster and Nakajima, 2012; Wu et al., 2016; Sennrich et al., 2016; Kudo and Richardson, 2018).

These subword unit approaches are linguistic-free and are mainly models, which are estimated on raw text. On the other side, it has been observed that these subword models do not correspond to linguistic units such as morphemes, affixes, etc. (Huck et al., 2017; Macháček et al., 2018; Ataman et al., 2017; Pinnis et al., 2017).

In order to conduct our subword comparison, we create three versions of the same ALBERT model, which are trained on the same data but with different tokenization. The goal of this subword study is to verify the impacts of subword models associated to our ALBERT models in NLP tasks. Especially in tokens class classification tasks, such as Named Entity Recognition or Spoken Language Understanding tasks.

Contributions: First, this paper presents the release of a multilingual version of ALBERT (Lan et al., 2019): mALBERT[1], trained on open-source and ethical data; second, we propose a study of the subword tokenization process focused on the vocabulary size impact. We also measure the tokenization impact, which is correlated with the subwords segmentation rate of tokens.

[1] https://huggingface.co/XXXXXXXX

LREC-COLING 2024, pages 11023–11029

The paper is organized as follows: first, we present the model architecture and the pre-training details in Section 2; Section 3 details the experiments conducted using our new models and the tokenization study; Finally, the last section presents the conclusion and outcomes of this paper.

2. Model Pre-training

As far as we know, there is no multilingual compact model. We therefore propose to pre-train a new version of ALBERT from scratch: *mALBERT*.

ALBERT is based on parameter sharing/reduction techniques that enable us to reduce the computational complexity and speed up training and inference phases. Compared to previous compact models such as DistilBERT (Sanh et al., 2019), Q-BERT (Shen et al., 2020) or TernaryBERT (Zhang et al., 2020), ALBERT is to the date the smallest pre-trained models with 12 million parameters and <50 megabyte model size. ALBERT models also show their ecological advantages regarding bigger models (Cattan et al., 2022).

2.1. Data

Aiming to use open-source and ethical data to pre-train the mALBERT model, we decided to use only Wikipedia data for each language. Figure 1 presents the language distribution of the Wikipedia corpus collected on January 2023. The corpus is roughly 21 billion words across 50 most common languages on Wikipedia, plus English and Basque.

As for many other multilingual models, English prevails the whole corpus with French, German, and Spanish. These four languages represent nearly 50% of the corpus.

2.2. Subword unit

The subword unit tokenization model chosen for our multilingual ALBERT model is based on a unigram language model approach (Kudo and Richardson, 2018). This subword unit approach was chosen because it enables us to fix the final amount of vocabulary.

Three subword unit models were trained on a subpart of the corpus selected randomly, in order to study the impact of the tokenization process on the final ALBERT model performances. The tokenization models differ only with the amounts of the final vocabulary generated: 32k, 64k, and 128k.

2.3. Training parameters

Models are trained for roughly 9000 hours on the ANONYMIZED CALCULATOR NAME, using the UER-py toolkit (Zhao et al., 2019) jointly with DeepSpeed (Rasley et al., 2020), and use multiple training objectives (masked language modeling and next sentence or sentence order prediction). We use the same learning configuration as the original model with a batch size of 128 and an initial learning rate set to 3.125×10^{-4}.

Finally, we pre-train three models based on the same amount of corpus, with the same amount of parameters, but they differ only by the amount of input vocabulary. We noted our final models as follows: *mALBERT-128k*, *mALBERT-64k*, and *mALBERT-32k*, which respectively use an amount of 128k, 64k, and 32k tokens.

3. Experiments

Our three new models are benchmarked on two kinds of classical NLP tasks: the slot-filling and classification tasks. These tasks use standard fine-tuning approaches, in which fine-tuning and evaluation scripts are provided by HuggingFace (Wolf et al., 2019). For each experiment, we do not seek to have the best score, but a point of comparison for our models.

We compare our new multilingual ALBERT model to the large multilingual model *mBERT* (Devlin et al., 2019) as well as on the compact multilingual models with a distilled version of mBERT: *distil-mBERT* (Sanh et al., 2019). Our comparison includes also some monolingual versions of ALBERT for English (noted *EnALBERT* in CoNLL2003 and MultiCoNER tasks) and French (in MEDIA), noted *FrALBERT* (Cattan et al., 2021).

Finally, we do not compare our models with bigger LLM such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023) or BLOOM (Scao et al., 2022), for resources and ecological considerations.

3.1. Slot-filling benchmark

Six slot-filling tasks are used to benchmark our new mALBERT models: two multilingual understanding tasks, Massively Multilingual NLU 2022 (MMNLU) (FitzGerald et al., 2022) and MultiATIS++ (Xu et al., 2020); two Named Entity Recognition monolingual tasks: CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) and MultiCoNER (Malmasi et al., 2022); and two monolingual language understanding tasks: SNIPS (Coucke et al., 2018) and MEDIA (Bonneau-Maynard et al., 2009).

Table 1 presents the results obtained on the slot-filling tasks according to the F1-measure. For every task and model, we perform 10 runs with a different seed each time. Over all tasks, *mBERT* and *Distil-mBERT* obtain the best results. On the one hand, monolingual ALBERT models perform better on CoNLL2003 and SNIPS tasks. On the other hand,

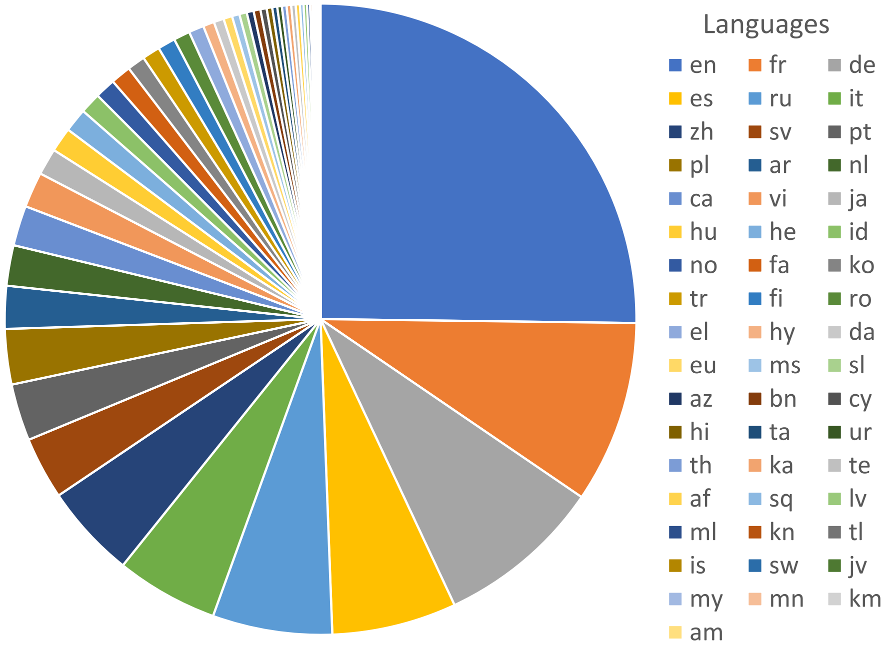

Figure 1: Language distribution (52 languages) over the training corpus. In the legend, languages are presented according to their representativity: from left to right and from up and down. The most representative language is English (*en*) and the least one is Amharic (*am*)

Models ╲ Tasks	MMNLU	MultiATIS++	CoNLL2003	MultiCoNER	SNIPS	MEDIA
mBERT	73.46* (0.11)	92.22 (0.11)	95.59* (0.10)	66.36* (0.18)	96.09 (0.31)	87.90* (0.09)
Distil-mBERT	72.44* (0.08)	91.69 (0.09)	94.59* (0.13)	61.26* (0.13)	94.95 (0.22)	86.83* (0.21)
EnALBERT	N/A	N/A	89.67* (0.34)	42.36* (0.22)	95.95 (0.13)	N/A
FrALBERT	N/A	N/A	N/A	N/A	N/A	81.76 (0.59)
mALBERT-128k	65.81* (0.11)	89.14 (0.15)	88.27* (0.24)	46.01* (0.18)	91.60 (0.31)	83.15* (0.38)
mALBERT-64k	65.29* (0.14)	88.88 (0.14)	86.44* (0.37)	44.70* (0.27)	90.84 (0.47)	82.30 (0.19)
mALBERT-32k	64.83* (0.22)	88.60 (0.27)	84.96* (0.41)	44.13* (0.39)	89.89 (0.68)	82.04 (0.28)

Table 1: Results on several slot-filling tasks regarding the F1-measure score. The results are the mean of 10 different runs, and the standard deviation is noted between parenthesis. $*$: p-value < 0.05.

one can observe that *mALBERT* models perform better than *FrALBERT* and *EnALBERT* models on MultiCoNER and MEDIA tasks, respectively. This ensure us that *mALBERT* is comparable with other monolingual ALBERT models.

In all slot-filling tasks, the 128k version of the mALBERT model performed better than the two other variants. Moreover, we observe a hierarchy in the mALBERT model versions according to their vocabulary size: the one with the smaller vocabulary is the worst and the 64k mALBERT variant is second.

3.2. Classification benchmark

For the classification benchmark, we evaluate our models against four tasks. First, two multilingual tasks: Massively Multilingual NLU 2022 (MMNLU) (FitzGerald et al., 2022) and MultiATIS++ (Xu et al., 2020) ; second, two monolingual tasks: SNIPS (Coucke et al., 2018) and Stanford Sentiment Treebank v2 (SST2) (Socher et al., 2013).

Like in the slot-filling task, bigger models obtained the best results over all tasks. Focusing on the mALBERT models, they obtained results with a $p-value$ lower than 0.05 only in the MMNLU task. On other tasks (MultiATIS++, SNIPS, and SST2), the significativity of results between our new models are not reached. Considering MMNLU results, this leads us to the same observation we have with the slot-filling task: the mALBERT model performances are ranked according to their vocabulary size.

3.3. Tokenization Impact

The starting point of this study is to measure the impact of the tokenization of subword unit models. We compare our three tokenization models: 32k, 64k, and 128k codes (which, in our case, corresponds to the final amount of vocabulary). This study focuses on a Named Entity task, the CoNLL2003 task, which is a slot-filling task based on a token classification method. This means the segmenta-

Models ╲ Tasks	MMNLU	MultiATIS++	SNIPS	SST2
mBERT	80.32* (0.09)	96.14* (0.17)	97.31 (0.31)	46.49* (0.76)
Distil-mBERT	78.23* (0.08)	92.79* (0.35)	97.69 (0.25)	43.59 (0.31)
EnALBERT	N/A	N/A	97.60 (0.11)	43.66 (1.88)
mALBERT-128k	72.35* (0.09)	90.58 (0.98)	96.84 (0.49)	34.66 (1.46)
mALBERT-64k	71.26* (0.11)	90.97 (0.70)	96.53 (0.44)	34.64 (1.02)
mALBERT-32k	70.76* (0.11)	90.55 (0.98)	96.49 (0.45)	34.18 (1.64)

Table 2: Results on several classification tasks regarding the Accuracy score. The results are the mean of 10 different runs, and the standard deviation is noted between parenthesis. *: p-value < 0.05.

Plain text	acquisition	of	Daniels	Pharmaceuticals	Inc	of	St.	Petersburg	,	Fla.
Reference	O	O	B-ORG	I-ORG	I-ORG	O	B-LOC	I-LOC	O	B-LOC
Tok-32k	_acquis␣i␣tion	_of	_Daniel␣s	_Pharmac␣e␣u␣tical␣s	_Inc	_of	_St␣.	_Petersburg	_,	_F␣la␣.
mALBERT-32k	O	O	B-ORG	I-ORG	B-ORG	I-ORG	I-ORG	I-ORG	O	B-ORG
Tok-64k	_acquisition	_of	_Daniel␣s	_Pharmac␣e␣u␣tical␣s	_Inc	_of	_St␣.	_Petersburg	_,	_Fla␣.
mALBERT-64k	O	O	B-PER	I-PER	B-ORG	O	B-ORG	I-ORG	O	B-PER
Tok-128k	_acquisition	_of	_Daniel␣s	_Pharmaceutical␣s	_Inc	_of	_St␣.	_Petersburg	_,	_Fla␣.
mALBERT-128k	O	O	B-ORG	I-ORG	I-ORG	O	B-LOC	I-LOC	O	B-LOC

Table 3: Example of segmentation / tokenization for each model and the label detected by the model for the CoNLL2003 task (NER). In this table the original input text is noted *Plain text*, with its gold labelization (*Reference*). Then each next row corresponds to a tokenization model (*Tok-32k*, *Tok-64k*, *Tok-128k*) and the output of the associated model (*mALBERT-32k*, *mALBERT-64k*, *mALBERT-128k*). The token segmentation in subwords is indicated with a special character as separator (␣).

Subword vocab. size	Tok-32k	Tok-64k	Tok-128k
NE	120.59 %	85.28 %	62.69 %
Not NE	57.64 %	36.04 %	25.37 %

Table 4: Impact of the tokenization on word type (i.e.: belong to a Named Entity or not.) in the CoNLL2003 task. We reported the percentage of additional segmentation observed.

tion of the token in subwords could increase the sentence context, which may impact the final labelization result.

In order to measure the possible impact of the subword tokenization, we estimate the amount of additional segmentation according to Name Entity (NE) labels (table 4). We can observe a significant impact on the token segmentation associated with NE: the 128k subword model segmentation of tokens produces 62% more subwords, meanwhile, the 32k subword model produces 120% of additional subwords.

We push deeper into the analysis and estimate the Pearson correlation score between the segmentation of the word in subwords and the non-detection of the associated label of the original token. The correlation score is 0.44, which implies a moderate correlation of the tokenization impact on the labelization process. This means the more the entity is segmented, the less accurate the model is to identify the right entity.

Table 3 presents an example of the segmentation and labelization of the sequence “acquisition of Daniels Pharmaceuticals Inc of St. Petersburg, Fla.”. In this example, we can observe that the most split word is “Pharmaceuticals”. This sequence of subwords illustrates the impact, especially on the label of next word ‘Inc”. The impact can directly be observed on the label of words “Pharmaceuticals” and “Fla.”. The right labelization is obtained once the whole segment is the less splitted in subwords.

The subword tokenization seems to interfere with the labelization of these tokens made by the model. Finally, these remarks on subword tokenization seem obvious. Still, as far as we know, we have not found any study on the impact of tokenization on pre-trained language models. This first study shall be pushed further to precisely measure the impact of subword tokenization models on other tasks and domains.

4. Conclusion

This paper presents the first multilingual ALBERT model (mALBERT), pre-trained on Wikipedia dump in 52 languages. The model comes with three vocabulary size variants: 32k, 64k, and 128k. All variants were pre-trained on data extracted from 91 Go of Wikipedia dumps, which represents more than 21 billion words.

So, is a multilingual compact still worth it? Evaluations in classical NLP tasks (slot-filling and classification tasks) show the multilingual version of ALBERT has comparable results to the monolingual versions used in this paper. From an ecological and resource aspect, one model pre-training on GPU time took 9k hours, which is far from the million hours for the BLOOM LLM.

The tokenization study, focused on vocabulary size, gives some feedback about the importance of the impact of subword tokenization. The moderate correlation observed on a classical Named Entity task enables us to say the more you split tokens into subwords, the less the Entity is well detected.

In the next steps, the extension of the subword tokenization model study will investigate which kind of segmentation could be the best for Pre-trained Language Models on more NLP tasks.

The three versions of the model are freely available on huggingFace[2]

5. Acknowledgements

This paper was funded by the *Multiligual SLU for Contextual Question Answering (MuSCQA)* project from "France Relance" of French government funded by French National Research Agency (ANR), grant number: ANR-21-PRRD-0001-01. This work was performed using HPC resources from GENCI–IDRIS (Grant 2023-A0131013834).

6. Bibliographical References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331–342.

Oralie Cattan, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2022. Benchmarking transformers-based models on french spoken language understanding tasks. In *INTERSPEECH 2022*.

Oralie Cattan, Christophe Servan, and Sophie Rosset. 2021. On the usability of transformers-based models for a French question-answering task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 244–255, Held Online. INCOMA Ltd.

Rohan Chitnis and John DeNero. 2015. Variable-length word encodings for neural translation models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2088–2093.

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.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67.

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: System Demonstrations*, pages 66–71.

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. *CoRR*, abs/1909.11942.

Dominik Macháček, Jonáš Vidra, and Ondřej Bojar. 2018. Morphological and language-agnostic word segmentation for nmt. In *International Conference on Text, Speech, and Dialogue*, pages 277–284. Springer.

Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf, editors. 2020. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Online.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pages 237–245. Springer.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

[2] https://huggingface.co/cservan/malbert-base-cased-32k
https://huggingface.co/cservan/malbert-base-cased-64k
https://huggingface.co/cservan/malbert-base-cased-128k

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-BERT: hessian based ultra low precision quantization of BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8815–8821. AAAI Press.

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, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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*.

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020. Ternarybert: Distillation-aware ultra-low bit bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 509–521.

Zhe Zhao, Hui Chen, Jinbin Zhang, Wayne Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. 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): System Demonstrations*, pages 241–246.

7. Language Resource References

Bonneau-Maynard, Hélene and Quignard, Matthieu and Denis, Alexandre. 2009. *MEDIA: a semantically annotated corpus of task oriented dialogs in French: Results of the French media evaluation campaign*. Springer.

Coucke, Alice and Saade, Alaa and Ball, Adrien and Bluche, Théodore and Caulier, Alexandre and Leroy, David and Doumouro, Clément and Gisselbrecht, Thibault and Caltagirone, Francesco and Lavril, Thibaut and others. 2018. *Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces*.

FitzGerald, Jack and Hench, Christopher and Peris, Charith and Mackie, Scott and Rottmann, Kay and Sanchez, Ana and Nash, Aaron and Urbach, Liam and Kakarala, Vishesh and Singh, Richa and others. 2022. *Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages*.

Malmasi, Shervin and Fang, Anjie and Fetahu, Besnik and Kar, Sudipta and Rokhlenko, Oleg.

2022. *SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER)*. Association for Computational Linguistics.

Rajpurkar, Pranav and Zhang, Jian and Lopyrev, Konstantin and Liang, Percy. 2016. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*.

Socher, Richard and Perelygin, Alex and Wu, Jean and Chuang, Jason and Manning, Christopher D and Ng, Andrew Y and Potts, Christopher. 2013. *Recursive deep models for semantic compositionality over a sentiment treebank*.

Tjong Kim Sang, Erik F. and De Meulder, Fien. 2003. *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*.

Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel. 2018. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*.

Xu, Weijia and Haider, Batool and Mansour, Saab. 2020. *End-to-End Slot Alignment and Recognition for Cross-Lingual NLU*. Association for Computational Linguistics.