Evaluating Contextualized Language Models for Hungarian

Judit Ács[1,2], Dániel Lévai[3], Dávid Márk Nemeskey[2], András Kornai[2]

[1] Department of Automation and Applied Informatics
Budapest University of Technology and Economics
[2] Institute for Computer Science and Control
[3] Alfréd Rényi Institute of Mathematics

Abstract. We present an extended comparison of contextualized language models for Hungarian. We compare huBERT, a Hungarian model against 4 multilingual models including the multilingual BERT model. We evaluate these models through three tasks, morphological probing, POS tagging and NER. We find that huBERT works better than the other models, often by a large margin, particularly near the global optimum (typically at the middle layers). We also find that huBERT tends to generate fewer subwords for one word and that using the last subword for token-level tasks is generally a better choice than using the first one.
Keywords: huBERT, BERT, evaluation

1   Introduction

Contextualized language models such BERT (Devlin et al., 2019) drastically improved the state of the art for a multitude of natural language processing applications. Devlin et al. (2019) originally released 4 English and 2 multilingual pretrained versions of BERT (mBERT for short) that support over 100 languages including Hungarian. BERT was quickly followed by other large pretrained Transformer (Vaswani et al., 2017) based models such as RoBERTa (Liu et al., 2019b) and multilingual models with Hungarian support such as XLM-RoBERTa (Conneau et al., 2019). Huggingface released the Transformers library (Wolf et al., 2020), a PyTorch implementation of Transformer-based language models along with a repository for pretrained models from community contribution [1]. This list now contains over 1000 entries, many of which are domain- or language-specific models.

Despite the wealth of multilingual and language-specific models, most evaluation methods are limited to English, especially for the early models. Devlin et al. (2019) showed that the original mBERT outperformed existing models on the XNLI dataset (Conneau et al., 2018b). mBERT was further evaluated by Wu and Dredze (2019) for 5 tasks in 39 languages, which they later expanded to over 50 languages for part-of-speech tagging, named entity recognition and dependency parsing (Wu and Dredze, 2020).

---

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

Nemeskey (2020) released the first BERT model for Hungarian named *hu-BERT* trained on Webcorpus 2.0 (Nemeskey, 2020, ch. 4). It uses the same architecture as BERT base with 12 Transformer layers with 12 heads and 768 hidden dimension each with a total of 110M parameters. huBERT has a Word-Piece vocabulary with 30k subwords.

In this paper we focus on evaluation for the Hungarian language. We compare huBERT against multilingual models using three tasks: morphological probing, POS tagging and NER. We show that huBERT outperforms all multilingual models, particularly in the lower layers, and often by a large margin. We also show that subword tokens generated by huBERT's tokenizer are closer to Hungarian morphemes than the ones generated by the other models.

2   Approach

We evaluate the models through three tasks: morphological probing, POS tagging and NER. Hungarian has a rich inflectional morphology and largely free word order. Morphology plays a key role in parsing Hungarian sentences.

We picked two token-level tasks, POS tagging and NER for assessing the sentence level behavior of the models. POS tagging is a common subtask of downstream NLP applications such as dependency parsing, named entity recognition and building knowledge graphs. Named entity recognition is indispensable for various high level semantic applications.

2.1   Morphological probing

Probing is a popular evaluation method for black box models. Our approach is illustrated in Figure 1. The input of a probing classifier is a sentence and a target position (a token in the sentence). We feed the sentence to the contextualized model and extract the representation corresponding to the target token. We use either a single Transformer layer of the model or the weighted average of all layers with learned weights. We train a small classifier on top of this representation that predicts a morphological tag. We expose the classifier to a limited amount of training data (2000 training and 200 validation instances). If the classifier performs well on unseen data, we conclude that the representation includes said morphological information. We generate the data from the automatically tagged Webcorpus 2.0. The target words have no overlap between train, validation and test, and we limit class imbalance to 3-to-1 which resulted in filtering some rare values. The list of tasks we were able to generate is summarized in Table 1.

2.2   Sequence tagging tasks

Our setup for the two sequence tagging tasks is similar to that of the morphological probes except we train a shared classifier on top of all token representations. Since multiple subwords may correspond to a single token (see Section 3.1 for
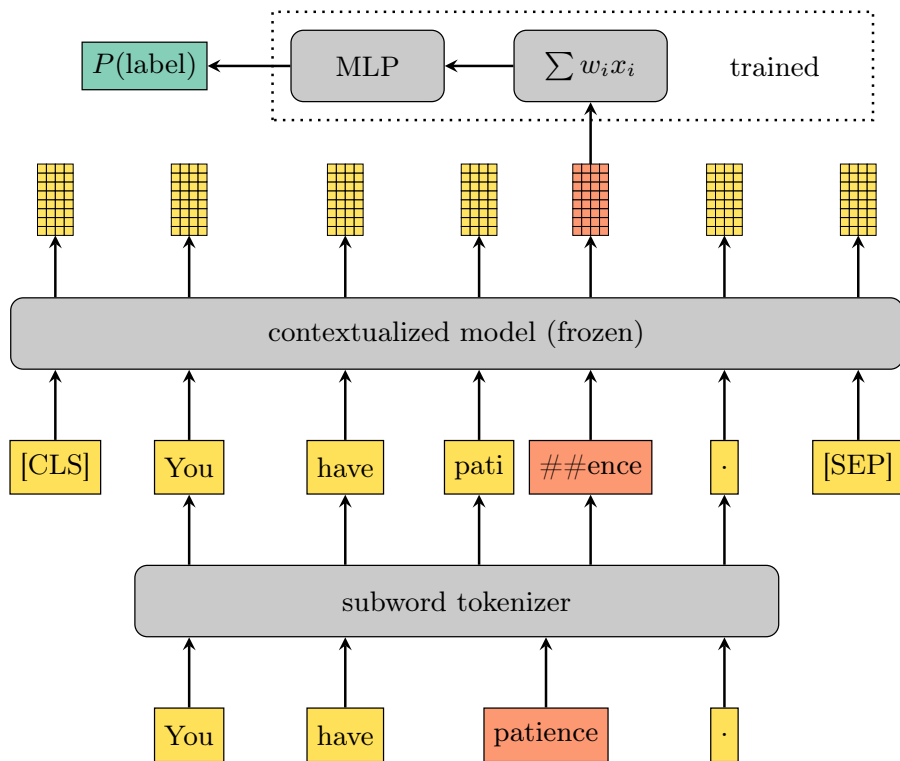
Fig. 1: Probing architecture. Input is tokenized into subwords and a weighted average of the mBERT layers taken on the last subword of the target word is used for classification by an MLP. Only the MLP parameters and the layer weights w_i are trained.

more details), we need to aggregate them in some manner: we pick either the first one or the last one.[2]

We use two datasets for POS tagging. One is the Szeged Universal Dependencies Treebank (Farkas et al., 2012; Nivre et al., 2018) consisting of 910 train, 441 validation, and 449 test sentences. Our second dataset is a subsample of Webcorpus 2 tagged with emtsv (Indig et al., 2019) with 10,000 train, 2000 validation, and 2000 test sentences.

Our architecture for NER is identical to the POS tagging setup. We train it on the Szeged NER corpus consisting of 8172 train, 503 validation, and 900 test sentences.

---

[2] We also experimented with other pooling methods such as elementwise max and sum but they did not make a significant difference.

Morph tag	POS	#classes	Values
Case	noun	18	Abl, Acc, . . . , Ter, Tra
Degree	adj	3	Cmp, Pos, Sup
Mood	verb	4	Cnd, Imp, Ind, Pot
Number psor	noun	2	Sing, Plur
Number	adj	2	Sing, Plur
Number	noun	2	Sing, Plur
Number	verb	2	Sing, Plur
Person psor	noun	3	1, 2, 3
Person	verb	3	1, 2, 3
Tense	verb	2	Pres, Past
VerbForm	verb	2	Inf, Fin

Table 1. List of morphological probing tasks.

2.3   Training details

We train all classifiers with identical hyperparameters. The classifiers have one hidden layer with 50 neurons and ReLU activation. The input and the output layers are determined by the choice of language model and the number of target labels. This results in 40k to 60k trained parameters, far fewer than the number of parameters in any of the language models.

All models are trained using the Adam optimizer (Kingma and Ba, 2014) with $lr = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$. We use 0.2 dropout for regularization and early stopping based on the development set.

3   The models evaluated

We evaluate 5 models.

huBERT the Hungarian BERT, is a BERT-base model with 12 Transformer layers, 12 attention heads, each with 768 hidden dimensions and a total of 110 million parameters. It was trained on Webcorpus 2.0 (Nemeskey, 2020), 9-billion-token corpus compiled from the Hungarian subset of Common Crawl[3]. Its string identifier in Huggingface Transformers is `SZTAKI-HLT/hubert-base-cc`.

mBERT the cased version of the multilingual BERT. It is a BERT-base model with identical architecture to huBERT. It was trained on the Wikipedias of the 104 largest Wikipedia languages. Its string id is `bert-base-multilingual-cased`.

XLM-RoBERTa the multilingual version of RoBERTa. Architecturally, it is identical to BERT; the only difference lies in the training regimen. XLM-RoBERTa was trained on 2TB of Common Crawl data, and it supports 100 languages. Its string id is `xlm-roberta-base`.

---

[3] `https://commoncrawl.org/`

XLM-MLM-100 is a larger variant of XLM-RoBERTa with 16 instead of 12 layers. Its string id is `xlm-mlm-100-1280`.

distilbert-base-multilingual-cased is a *distilled* version of mBERT. It cuts the parameter budget and inference time by roughly 40% while retaining 97% of the tutor model's NLU capabilities. Its string id is `distilbert-base-multilingual-cased`.

3.1   Subword tokenization

Subword tokenization is a key component in achieving good performance on morphologically rich languages. Out of the 5 models we compare, huBERT, mBERT and DistilBERT use the WordPiece algorithm (Schuster and Nakajima, 2012), XLM-RoBERTa and XLM-MLM-100 use the SentencePiece algorithm (Kudo and Richardson, 2018). The two types of tokenizers are algorithmically very similar, the differences between the tokenizers are mainly dependent on the vocabulary size per language. The multilingual models consist of about 100 languages, and the vocabularies per language are (not linearly) proportional to the amount of training data available per language. Since huBERT is trained on monolingual data, it can retain less frequent subwords in its vocabulary, while mBERT, RoBERTa and MLM-100, being multilingual models, have token information from many languages, so we anticipate that huBERT is more faithful to Hungarian morphology. DistilBERT uses the tokenizer of mBERT, thus it is not included in this subsection.

	huBERT	mBERT	RoBERTa	MLM-100	emtsv
Languages	1	104	100	100	1
Vocabulary size	32k	120k	250k	200k	–
Entropy of first WP	8.99	6.64	6.33	7.56	8.26
Entropy of last WP	6.82	6.38	5.60	6.89	5.14
More than one WP	94.9%	96.9%	96.5%	97.0%	95.8%
Length in WP	2.8±1.4	3.9±1.8	3.2±1.4	3.5±1.5	3.1±1.1
Length of first WP	4.3±3.0	2.7±1.9	3.5±2.7	3.1±2.0	5.2±2.4
Length of last WP	3.8±2.9	2.6±1.8	3.1±2.2	2.8±1.8	2.7±1.7
Accuracy to emtsv	0.16	0.05	0.14	0.08	1.00
Accuracy to emtsv in first WP	0.41	0.26	0.44	0.33	1.00
Accuracy to emtsv in last WP	0.43	0.41	0.47	0.39	1.00

Table 2. Measures on the train data of the POS tasks. The length of first and last WP is calculated in characters, while the word length is calculated in WPs. DistilBERT data is identical to mBERT.

As shown in Table 2, there is a gap between the Hungarian and multilingual models in almost every measure. mBERT's shared vocabulary consists only of 120k subwords for all 100 languages while huBERT's vocabulary contains 32k

items and is uniquely for Hungarian. Given the very limited inventory of mBERT, only the most frequent Hungarian words are represented as a single token, while longer Hungarian words are segmented, often very poorly. The average number of subwords a word is tokenized into is 2.77 in the case of huBERT, while all the other models have significantly higher mean length. This does not pose a problem in itself, since the tokenizers work with a given dictionary size and frequent words need not to be segmented into subwords. But in case of words with rarer subwords, the limits of smaller monolingual vocabulary can be observed, as shown in the following example: *szállítójárművekkel* 'with transport vehicles'; *szállító-jármű-vek-kel* 'transport-vehicle-PL-INS' for huBERT, *sz-ál-lí-tó-já-rm-ű-vek-kel* for mBERT, which found the affixes correctly (since affixes are high in frequency), but have not found the root 'transport vehicle'.
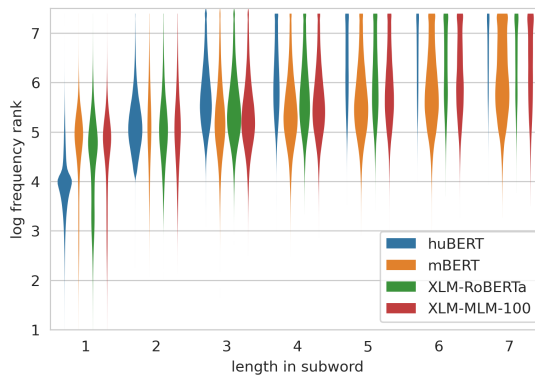

Fig. 2: Distribution of length in subword vs. log frequency rank. The count of words for one subword length is proportional to the size of the respective violin.

Distributionally, huBERT shows a stronger Zipfian distribution than any other model, as shown in Figure 2. Frequency and subword length are in a linear relationship for the huBERT model, while in case of the other models, the subword lengths does not seem to be correlated the log frequency rank. The area of the violins also show that words typically consist of more than 3 subwords for the multilingual models, contrary to the huBERT, which segments the words typically into one or two subwords.

4   Results

We find that huBERT outperforms all models in all tasks, often with a large margin, particularly in the lower layers. As for the choice of subword pooling (first or last) and the choice of layer, we note some trends in the following subsections.

4.1   Morphology

The last subword is always better than the first subword except for a few cases for degree ADJ. This is not surprising because superlative is marked with a circumfix and it is differentiated from comparative by a prefix. The rest of the results in this subsection all use the last subword.

huBERT is better than all models, especially in the lower layers in morphological tasks, as shown in Figure 3. However, this tendency starts at the second layer and the first layer does not usually outperform the other models. In some morphological tasks huBERT systematically outperforms the other models: these are mostly the simpler noun and adjective-based probes. In possessor tasks (tagged [psor] in Figure 3) XLM models are comparable to huBERT, while mBERT and distil-mBERT generally perform worse then huBERT. In verb tasks XLM-RoBERTa achieves similar accuracy to huBERT in the higher layers, while in the lower layers, huBERT tends to have a higher accuracy.

HuBERT is also better than all models in all tasks when we use the weighted average of all layers as illustrated by Figure 4. The only exceptions are adjective degrees and the possessor tasks. A possible explanation for the surprising effectiveness of XLM-MLM-100 is its higher layer count.

4.2   POS tagging

Figure 5 shows the accuracy of different models on the gold-standard Szeged UD and on the silver-standard data created with emtsv.

Last subword pooling always performs better than first subword pooling. As in the morphology tasks, the XLM models perform only a bit worse than huBERT. mBERT is very close in performance to huBERT, unlike in the morphological tasks, while distil-mBERT performs the worst, possibly due to its far lower parameter count.

We next examine the behavior of the layers by relative position.[4] The embedding layer is a static mapping of subwords to an embedding space with a simple positional encoding added. Contextual information is not available until the first layer. The highest layer is generally used as the input for downstream tasks. We also plot the performance of the middle layer. As Figure 6 shows, the embedding layer is the worst for each model and, somewhat surprisingly, adding one contextual layer only leads to a small improvement. The middle layer is actually better than the highest layer which confirms the findings of Tenney et al. (2019a) that BERT rediscovers the NLP pipeline along its layers, where POS tagging is a mid-level task. As for the choice of subword, the last one is generally better, but the gap shrinks as we go higher in layers.

4.3   Named entity recognition

In the NER task (Figure 7), all of the models perform very similarly in the higher layers, except for distil-mBERT which has nearly 3 times the error of

---

[4] We only do this on the smaller Szeged dataset due to resource limitations.
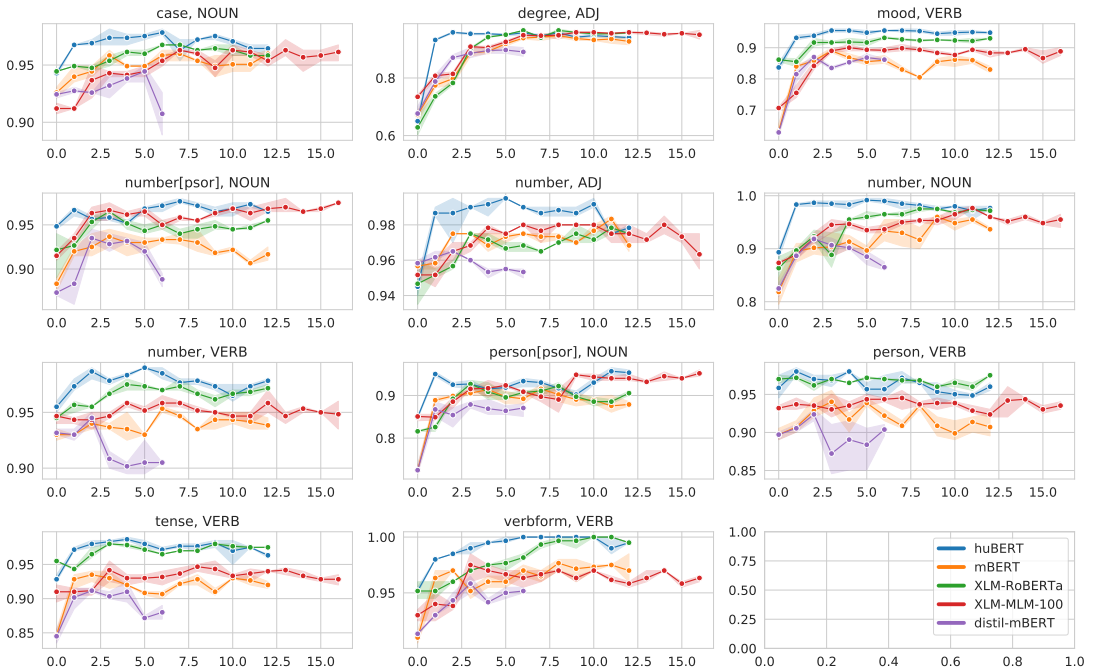
Fig. 3: The layerwise accuracy of morphological probes using the last subword. Shaded areas represent confidence intervals over 3 runs.
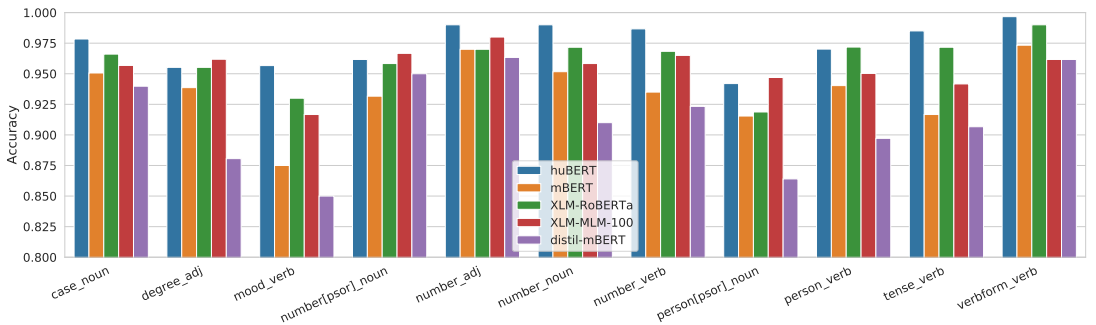

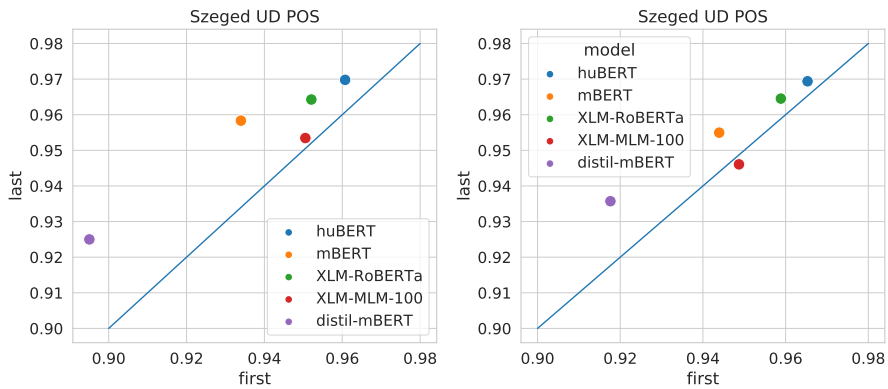Fig. 4: Probing accuracy using the weighted sum of all layers.

Fig. 5: POS tag accuracy on Szeged UD and on the Webcorpus 2.0 sample
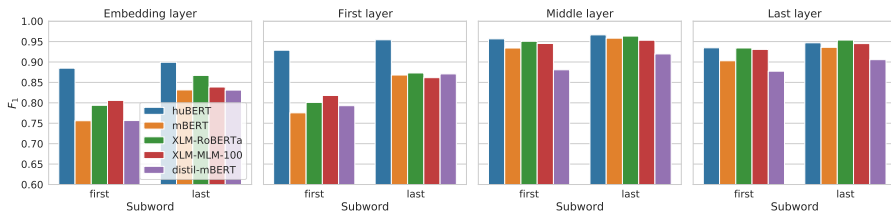

Fig. 6: Szeged POS at 4 layers: embedding layer, first Transformer layer, middle layer, and highest layer.
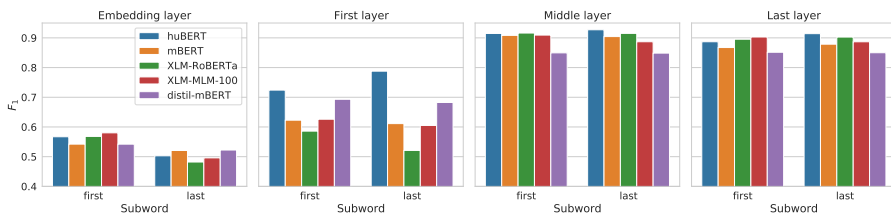

Fig. 7: NER F_1 score at the lowest, middle and highest layers.

the best model, huBERT. The closer we get to the global optimum, the clearer huBERT's superiority becomes. Far away from the optimum, when we use only the embedding layer, first subword is better than last, but the closer we get to the optimum (middle and last layer), the clearer the superiority of the last subword choice becomes.

5   Related work

Probing is a popular method for exploring blackbox models. Shi et al. (2016) was perhaps the first one to apply probing classifiers to probe the syntactic knowledge of neural machine translation models. Belinkov et al. (2017) probed NMT models for morphology. This work was followed by a large number of similar probing papers (Belinkov et al., 2017; Adi et al., 2017; Hewitt and Manning, 2019; Liu et al., 2019a; Tenney et al., 2019b; Warstadt et al., 2019; Conneau et al., 2018a; Hupkes and Zuidema, 2018). Despite the popularity of probing classifiers, they have theoretical limitations as knowledge extractors (Voita and Titov, 2020), and low quality of silver data can also limit applicability of important probing techniques such as canonical correlation analysis (Singh et al., 2019),

Multilingual BERT has been applied to a variety of multilingual tasks such as dependency parsing (Kondratyuk and Straka, 2019) or constituency parsingKitaev et al. (2019). mBERT's multilingual capabilities have been explored for NER, POS and dependency parsing in dozens of language by Wu and Dredze (2019) and Wu and Dredze (2020). The surprisingly effective multilinguality of mBERT was further explored by Dufter and Schütze (2020).

6   Conclusion

We presented a comparison of contextualized language models for Hungarian. We evaluated huBERT against 4 multilingual models across three tasks, morphological probing, POS tagging and NER. We found that huBERT is almost always better at all tasks, especially in the layers where the optima are reached. We also found that the subword tokenizer of huBERT matches Hungarian morphological segmentation much more faithfully than those of the multilingual models. We also show that the choice of subword also matters. The last subword is much better for all three kinds of tasks, except for cases where discontinuous morphology is involved, as in circumfixes and infixed plural possessives (Antal, 1963; Mel'cuk, 1972). Our data, code and the full result tables are available at `https://github.com/juditacs/hubert_eval`.

Acknowledgements

Bibliography

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., Goldberg, Y.: Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In: Proceedings of International Conference on Learning Representations (2017)

Antal, L.: The possessive form of the Hungarian noun. Linguistics 3, 50–61 (1963)

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., Glass, J.: What do neural machine translation models learn about morphology? In: Proc. of ACL (2017), `https://www.aclweb.org/anthology/P17-1080`

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2019)

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2126–2136. Association for Computational Linguistics (2018a), `http://aclweb.org/anthology/P18-1198`

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2018b)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (6 2019), `https://www.aclweb.org/anthology/N19-1423`

Dufter, P., Schütze, H.: Identifying elements essential for BERT's multilinguality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4423–4437. Association for Computational Linguistics, Online (11 2020), `https://www.aclweb.org/anthology/2020.emnlp-main.358`

Farkas, R., Vincze, V., Schmid, H.: Dependency parsing of Hungarian: Baseline results and challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 55–65. EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), `http://dl.acm.org/citation.cfm?id=2380816.2380826`

Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. 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). pp. 4129–4138 (2019)

Hupkes, D., Zuidema, W.: Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. In: Proc. of IJCAI (2018), `https://doi.org/10.24963/ijcai.2018/796`

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – Egy formátum mind felett [emtsv – One format to rule them all]. In: Berend, G., Gosztolya, G., Vincze, V. (eds.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2019)

Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014), `https://arxiv.org/abs/1412.6980`

Kitaev, N., Cao, S., Klein, D.: Multilingual constituency parsing with self-attention and pre-training. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3499–3505. Association for Computational Linguistics, Florence, Italy (7 2019), `https://www.aclweb.org/anthology/P19-1340`

Kondratyuk, D., Straka, M.: 75 languages, 1 model: Parsing universal dependencies universally. 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). pp. 2779–2795. Association for Computational Linguistics, Hong Kong, China (11 2019), `https://www.aclweb.org/anthology/D19-1279`

Kudo, T., Richardson, J.: 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. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (11 2018), `https://www.aclweb.org/anthology/D18-2012`

Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., Smith, N.A.: Linguistic knowledge and transferability of contextual representations pp. 1073–1094 (2019a), `https://www.aclweb.org/anthology/N19-1112`

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach (2019b)

Mel'cuk, I.A.: On the possessive forms of the Hungarian noun. In: Kiefer, F., Rouwet, N. (eds.) Generative grammar in Europe, pp. 315–332. Reidel, Dordrecht (1972)

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)

Nivre, J., Abrams, M., Agić, Ž., et al.: Universal Dependencies 2.3 (2018), `http://hdl.handle.net/11234/1-2895`, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

Schuster, M., Nakajima, K.: Japanese and Korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5149–5152. IEEE (2012)

Shi, X., Padhi, I., Knight, K.: Does string-based neural MT learn source syntax? In: Proc. of EMNLP (2016), `https://www.aclweb.org/anthology/D16-1159`

Singh, J., McCann, B., Socher, R., Xiong, C.: BERT is not an interlingua and the bias of tokenization. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). pp. 47–55. Association for Computational Linguistics, Hong Kong, China (11 2019), `https://www.aclweb.org/anthology/D19-6106`

Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (7 2019a), `https://www.aclweb.org/anthology/P19-1452`

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S., Das, D., Pavlick, E.: What do you learn from context? Probing for sentence structure in contextualized word representations. In: Proc. of ICLR (2019b), `https://openreview.net/forum?id=SJzSgnRcKX`

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`

Voita, E., Titov, I.: Information-theoretic probing with minimum description length. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 183–196. Association for Computational Linguistics, Online (11 2020), `https://www.aclweb.org/anthology/2020.emnlp-main.14`

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.F., Phang, J., Mohananey, A., Htut, P.M., Jeretic, P., Bowman, S.R.: Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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). pp. 2877–2887. Association for Computational Linguistics, Hong Kong, China (11 2019), `https://www.aclweb.org/anthology/D19-1286`

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), `https://www.aclweb.org/anthology/2020.emnlp-demos.6`

Wu, S., Dredze, M.: 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). pp. 833–844. Association for Computational Linguistics, Hong Kong, China (11 2019), `https://www.aclweb.org/anthology/D19-1077`

Wu, S., Dredze, M.: Are all languages created equal in multilingual BERT? In: Proceedings of the 5th Workshop on Representation Learning for NLP. pp. 120–130. Association for Computational Linguistics, Online (7 2020), `https://www.aclweb.org/anthology/2020.repl4nlp-1.16`