Multilingual BERT learns abstract case representations

Isabel Papadimitriou
Stanford University
isabelvp@stanford.edu

Ethan Chi
Stanford University
ethanchi@cs.stanford.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Kyle Mahowald
University of California, Santa Barbara
mahowald@ucsb.edu

1 Introduction

A key component of sentence meaning is the grammatical relationships between constituent elements of a sentence, such as between a verb and its subject and object. These relationships are distinct from semantic roles, because different languages map different semantic roles onto different grammatical relationships. Here we study whether modern neural embeddings represent these higher-order, abstract grammatical relationships by studying case transfer and generalization across languages with different morphosyntactic alignments.

In transitive sentences, languages need a way to specify which argument is the subject (A) and which is the object (O). In English, this is marked by word order ("The dog chased the lawyer"); in other languages, through grammatical case.

Most such languages have *nominative–accusative* case systems, where intransitive subjects (S) and transitive subjects (A) pattern together and are both marked with nominative case, whereas objects of transitive sentences (O) are marked with accusative case. The most common alternative is the *ergative–absolutive* system (e.g., in Basque), which marks transitive subjects as ergative, whereas both intransitive subjects and transitive objects share absolutive case (Figure 1).

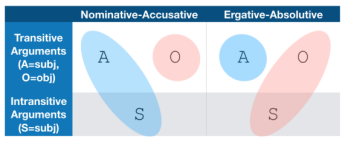

Figure 1: Illustration of the two predominant morphosyntactic alignment systems in world languages (Payne, 1997).

Recent work has demonstrated that deep neural models of language, such as BERT (Devlin et al., 2019), encode sentences in structurally meaningful ways (Manning et al., 2020; Linzen et al., 2016; Wilcox et al., 2018). However, these studies still leave an open question: are higher-order abstract grammatical features such as morphosyntactic alignment accessible to deep neural models?

To explore these questions, we study a large multilingual model (Multilingual BERT; henceforth mBERT) and take advantage of the typological differences in the way that languages represent subject and object to investigate whether mBERT represents case and grammatical role information separately from semantic role, and how it manages those representations across languages with different morphosyntactic alignment systems.

2 Method

For each of the 14 cased languages in mBERT that have a sufficiently large Universal Dependencies corpus (Nivre et al., 2016), we train a 2-layer case classifier that predicts the case of a noun (labelled using the UD corpus) from the BERT hidden state representing that noun. We only train the classifier on transitive subjects and objects (A and O nouns), leaving intransitive subjects (S nouns) out-of-domain. We train different classifiers for each of the 13 layers of BERT embeddings, in order to be able to detect any layer effects.

We test each classifier trained on language L_0 on four domains: a) transitive subjects and objects in L_0 (similar to those seen in training, to obtain an out-of-sample accuracy for each classifier), b) intransitive subjects in L_0 (a type of argument never seen in training, which showcase what morphosyntactic alignment the classifier has deduced from the training data), c) transitive subjects and objects in a new different language L_1 (which may or may not be a case-marked lan-

guage), and d) intransitive subjects in a new language L_1. See Figure 2 for a visualization of the classifier. We do this over every possible $(casedlanguage, language)$ pair in our training set, for each BERT layer.

In our analysis, we particularly focus on comparing the transfer performance of the case classifier between languages that have different morphosyntactic alignment systems (e.g., transfer from an ergative–absolutive language like Basque compared to transfer from a language with a nominative–accusative system like Greek).

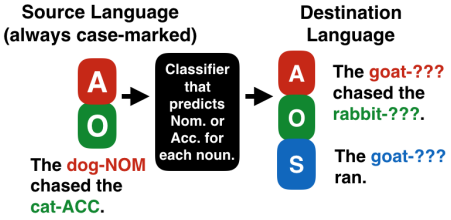

Figure 2: Illustration of the training and test process. We train the model to predict case on a cased language (e.g. Greek, Finnish, Russian) and test on either a cased or uncased language. We hold out intransitive subjects at training time, but not at test time.

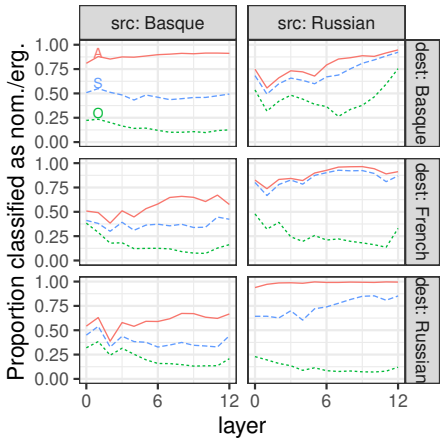

Figure 3: Across layers (x-axis), the proportion of the time that each of the A, S, and O arguments is classified as nominative or ergative. When the source language is Basque (left column), S is less likely to pattern with A. The full matrix is 14 cased source languages by 28 destination languages.

3 Results

1) Our case classifier can transfer zero-shot across languages with relatively high accuracy, suggesting that case and subjecthood relations are encoded in parallel ways across languages.

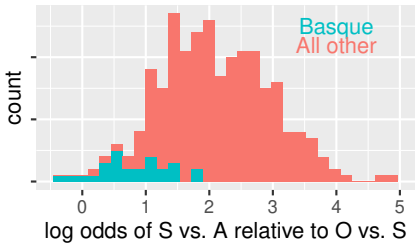

Figure 4: Log odds ratio representing how close S falls to A (relative to O) across languages. Experiments where the case classifier was trained on Basque appear in blue, and nominative-accusative languages in red. Basque is clearly clustered on the left.

2) When trained on transitive sentences in Basque (an ergative–absolutive language), mBERT's classification of intransitive subjects across other languages is an outlier relative to its behavior when trained on nominative–accusative languages. This suggests that BERT encodes representations that include the higher level grammatical features of morphosyntactic alignment even in transitive sentences where this is not realized.

3) mBERT shows an overall nominative–accusative bias, such that held-out intransitive subjects are classified with transitive subjects for most languages. This is in keeping with Ravfogel et al. (2019)'s finding that case generalization is harder than expected in Basque.

4) In an error analysis of a subject classifier trained on English, we find that mBERT more confidently characterizes animate nouns with high imageability and agency as subjects. In a further analysis on Czech, we find a strong correlation between the probability an S argument was labeled nominative and its animacy, suggesting that mBERT may represent grammatical subjecthood in a gradient way, such that high-animacy nouns are more "subject-y". This is consistent with characterizations of subjecthood in the functional linguistic tradition (Croft, 2001; Comrie, 1989).

4 Conclusion

Taken together, these results point the way towards further work characterizing the nature of grammatical role and and abstract case in deep neural models of language.

References

Bernard Comrie. 1989. *Language Universals and Linguistic Typology*, 2nd edition. University of Chicago Press, Chicago.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

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.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Thomas Edward Payne. 1997. *Describing morphosyntax: A guide for field linguists*. Cambridge University Press.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. 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 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of BlackboxNLP*, Brussels.