Proceedings of the Society for Computation in Linguistics

Volume 4 Article 51

2021

Multilingual BERT, Ergativity, and Grammatical Subjecthood

Isabel Papadimitriou
Stanford University, isabelvp@stanford.edu

Ethan A. Chi
Stanford University, ethanchi@cs.stanford.edu

Richard Futrell
rfutrell@uci.edu

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

Recommended Citation

Papadimitriou, Isabel; Chi, Ethan A.; Futrell, Richard; and Mahowald, Kyle (2021) "Multilingual BERT, Ergativity, and Grammatical Subjecthood," *Proceedings of the Society for Computation in Linguistics*: Vol. 4 , Article 51.
DOI: https://doi.org/10.7275/6p1w-g122
Available at: https://scholarworks.umass.edu/scil/vol4/iss1/51

Multilingual BERT, ergativity, and grammatical subjecthood

Isabel Papadimitriou
Stanford University
isabelvp@stanford.edu

Ethan A. 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

We investigate how Multilingual BERT (mBERT) encodes grammar by examining how the high-order grammatical feature of *morphosyntactic alignment* (how different languages define what counts as a "subject") is manifested across the embedding spaces of different languages.

Continuing a line of inquiry into how deep neural models process language (Manning et al., 2020; Linzen et al., 2016), our goal is to understand whether, and how, large pretrained models encode abstract features of the grammars of languages. To do so, we analyze the notion of subjecthood in Multilingual BERT (mBERT) across diverse languages with different **morphosyntactic alignments**. Alignment is a feature of the grammar of a language, rather than of any single word or sentence, letting us analyze mBERT's representation of language-specific high-order grammatical properties.

For 24 languages, we train small classifiers to distinguish the mBERT embeddings of nouns that are subjects of transitive sentences from nouns that are objects. We then test these classifiers on out-of-domain examples *within* and *across* languages. We go beyond standard probing methods (which rely on classifier accuracy to make claims about embedding spaces) by (a) testing the classifiers out-of-domain to gain insights about the shape and characteristics of the subjecthood classification boundary and (b) testing for awareness of morphosyntactic alignment, which is a feature of the grammar rather than of the classifier inputs.

In Experiment 1, we test our subjecthood classifiers on out-of-domain *intransitive subjects* (subjects of verbs which do not have objects, like "I slept") in their training language. Whereas in English and many other languages, we think of intransitive subjects as grammatical subjects, ergative languages have a different morphosyntactic alignment system that aligns intransitive subjects with objects (Dixon, 1979; Du Bois, 1987). We find evidence that a language's alignment is represented in mBERT's embeddings, as shown in Figure 2.

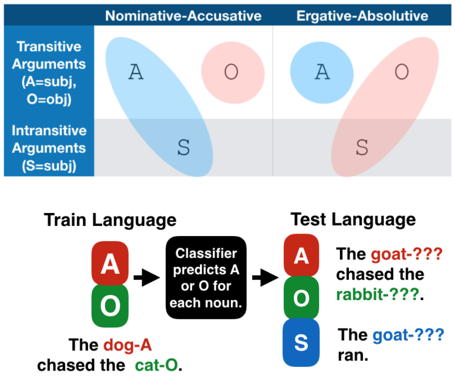
Nominative-Accusative
Ergative-Absolutive
Transitive Arguments (A=subj, O=obj)
Intransitive Arguments (S=subj)
A
O
S
A
O
S
Train Language
Test Language
A
O
Classifier predicts A or O for each noun.
The dog-A chased the cat-O.
A
O
S
The goat-??? chased the rabbit-???.
The goat-??? ran.

Figure 1: **Top**: Illustration of the difference between alignment systems. A (for agent) is notation used for the **transitive subject**, and O for the *transitive object*: "**The lawyer** chased *the dog*." S denotes the intransitive subject: "The lawyer laughed." The blue circle indicates which roles are marked as "subject" in each system. **Bottom**: Illustration of the training and test process. We train a classifier to distinguish A from O arguments using the BERT contextual embeddings, and test the classifier's behavior on intransitive subjects (S). The resulting distribution reveals to what extent morphosyntactic alignment (above) affects model behavior.

In Experiment 2, we perform successful zero-shot cross-linguistic transfer of our subject classifiers, finding that higher-order features of the grammar of each language are represented in a way that is parallel across languages. Zero-shot transfer of subjecthood classification is effective across languages. The average accuracy across all source-destination pairs for a high-performing mBERT layer (layer 10) is 82.61%. We can then look at how S is classified: does the subjecthood

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 425-426.
Held on-line February 14-19, 2021

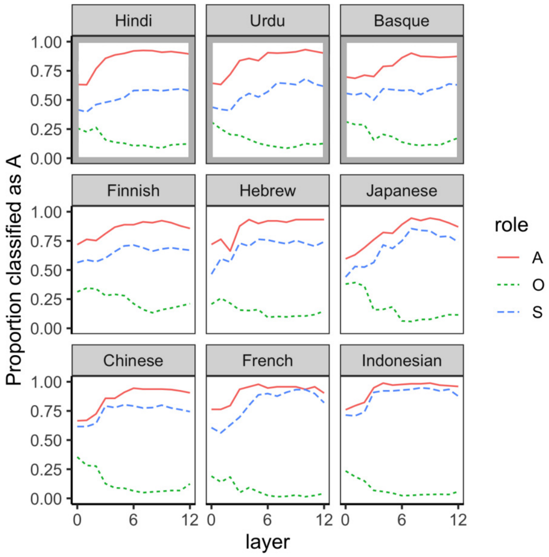

Figure 2: The behavior of subjecthood classifiers across mBERT layers (x-axis). For each layer, the proportion of the time that the classifier predicts arguments to be A, separated by grammatical role. In higher layers, A and O are reliably classified correctly, and S is mostly classified as A. When the source language is ergative or split-ergative (see gray outlined boxes), S is more intermediate between A and O.

of S, and the degree of ergativity within each language that we saw expressed in Experiment 1 generalize across languages? Classifiers trained on ergative languages are significantly more likely to classify S nouns in other languages as O (the source language's case system is a significant predictor of the probability of S being an agent, in a mixed effect regression with a random intercept for language $\beta = .11$, $t = 2.63$, $p < .05$) . Our results show that the ergative nature of these languages is encoded in the contextual embeddings of transitive nouns (where ergativity is not realized), and that this encoding of ergativity transfers coherently across languages.

In Experiment 3, we characterize the basis for these classifier decisions by studying how they vary as a function of linguistic features like passive constructions, animacy and grammatical case. We find that subjects which are passive are less likely to be categorized as subjects, as are subjects that are inanimate (as shown in Figure 3 or in less agentive cases (e.g., not nominative or ergative). We take this as evidence for a multifactored, probabilistic notion of subjecthood, as has been argued by Comrie (1981) and Hopper and Thompson (1980).

Taken together, the results of these experiments suggest that mBERT represents subjecthood and objecthood robustly and probabilistically. Its representation is general enough such that it can transfer across languages, but also language-specific enough that it learns language-specific abstract grammatical features.

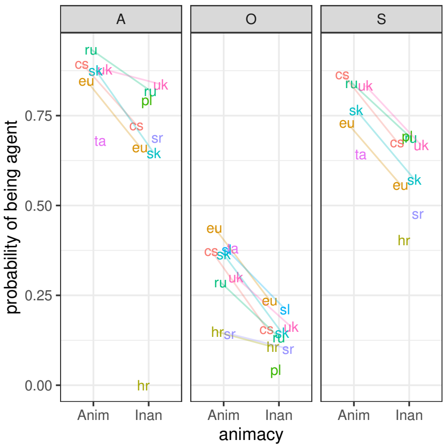

Figure 3: For a high-performing layer (Layer 10), the average probability of classifiers in all languages classifying nouns in languages with animacy distinctions as A. For all three roles, animates are more likely to be classified as agents. The labels are two-letter codes for the langauges.

Acknowledgments. We thank Dan Jurafsky, Jesse Mu, and Ben Newman for helpful comments. This work was supported by NSF grant #1947307 to R.F.

References

Bernard Comrie. 1981. *Language Universals and Linguistic Typology*, 1st edition. University of Chicago Press, Chicago.

Robert MW Dixon. 1979. Ergativity. *Language*, pages 59–138.

John W Du Bois. 1987. The discourse basis of ergativity. *Language*, pages 805–855.

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. *language*, pages 251–299.

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*.