Linguistic Complexity Through Form-Meaning Pairings: An Information Theoretical Approach to Equi-Complexity of Language

This research tackles the question: "Are all languages equally complex?" The research considers not only the formal aspects of language but also semantics, using an information-theoretical approach. Although the above question, referring to the equi-complexity of all languages, has been understood by linguists for a long time, the concept of equi-complicity still has to be verified. One of the main reasons the question has not been answered is that there is no common or overall method to measure the complexity of languages. This study seeks to measure the overall complexity of languages by measuring the unpredictability of what meaning a certain form represents in a certain context.

To compare the complexity of languages, computational linguists usually employ Shannon's entropy (Shannon, 1948). For instance, comparing the entropy rates and unigram entropy of words from more than 1,000 languages, Bentz et al. (2017) demonstrated that word entropies occupy a relatively narrow range, due to both word learnability and word expressivity. One of the limitations of this research is that it only considered formal aspects, but not semantics, because of the difficulty of using a computational approach for semantic aspects. In contrast to previous studies, this research considers a language as a set of form-meaning pairings, which requires dealing computationally with semantics. One of the major approaches for this is word embedding, in which the meanings of words are embedded in a multi-dimensional vector space based on computational processing.

This research estimates to how many meanings each form corresponds by clustering the embeddings. Based on that, the Shannon entropy, a metric quantifying the average unpredictability associated with discerning the intended meaning of a given form, is calculated using the formula:

$$H = \sum_{i=1}^{n} p(x_i) \log_2 p(x_i),$$

where H denotes the entropy of a form encompassing n distinct meanings, with $p(x_i)$ representing the probability of the ith meaning.

A pilot study compared the entropies of words from 10 languages (Chinese, Czech, English, French, German, Greek, Hebrew, Japanese, Russian and Spanish), using a pre-trained embedding model (Che et al., 2018), embeddings from language models (ELMo; Peters et al., 2018). The pilot study demonstrated that words in these 10 languages represent approximately one or two meanings on average, suggesting these languages have a similar distribution of form-meaning correspondence. In a follow-up survey, this research uses byte pair encoding (BPE; Gage, 1994) as a unit of form, to find a more applicable unit for every language.

References

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. Entropy, 19(6), 275. https://doi.org/10.3390/e19060275

Che, W., Liu, Y., Wang, Y., Zheng, B., & Liu, T. (2018). Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. Proceedings of The, 55–64. https://doi.org/10.18653/v1/K18-2005

Gage, P. (1994). A new algorithm for data compression. The C Users Journal archive, 12, 23-38.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2227–2237. https://doi.org/10.18653/v1/N18-1202

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x