Challenging datasets to test human's and LM linguistic knowledge

Proposers/Proposatzaileak: Itziar Gonzalez-Dios
Contact/Kontaktua: itziar.gonzalezd@ehu.eus

Description/Deskribapena:
Language models (LMs) trained in larger datasets have proven to be very powerful in identifying language patterns and reproducing them in different scenarios, but when it comes to assessing their linguistic competencies, they fail with functional words, negation, roles… (Kim et al., 2019; Ettinger, 2020; Ribeiro et al., 2020). Syntactic structures are also challenging (Linzen and Baroni, 2020).

On the other hand, there is an increase of effort to include gaze features (features derived from eye-tracking data) in LMs (Hollenstein et al., 2019, Barret and Hollenstein, 2020; Hollenstein et al., 2020).

In this work, we plan to create benchmarks that will be evaluated together with humans via eye-tracking experiments and with language models. The aim is to see how far/close are LMs responses to human abilities. Specifically, the data curation and evaluation of the LMs will be the scope of this project. The evaluation with humans will be carried out by other partners in the project.

There is a possibility of funding if the data is collected in Basque or Spanish, as this proposal is in line with the project ‘multilingual data collection.

Goals/Helburuak:
Creating a challenging benchmark (linguistic phenomena and tasks to be agreed with the student or according to the project’s needs) that to be tested with LMs.

Requirements/Betebeharrak:
- Good knowledge of the target language
- Basic programming skills (e.g. Python for NLP)
- Basic knowledge of language models (e.g. BERT)

Framework/Esparrua:

This study is framed in the current trend on analysing the LMs. We are encouraged to carry out this study in languages other than English.

Tasks and plan/Atazak eta plana:
- Selection of the knowledge and relations that will be the target of the experiment.
- Generation of the test dataset.
- Evaluation of the language model through the test dataset.

References/Erreferentziak:

Barrett, M., & Hollenstein, N. (2020). Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for Natural Language Processing. *Language and Linguistics Compass*, *14*(11), 1-16.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8, 34-48.

Hollenstein, N., Barrett, M., Troendle, M., Bigiolli, F., Langer, N., & Zhang, C. (2019). Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Hollenstein, N., Barrett, M., & Beinborn, L. (2020). Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources* (pp. 15-27).

Kim, N., Patel, R., Poliak, A., Xia, P., Wang, A., McCoy, T., ... & Pavlick, E. (2019). Probing What Different NLP Tasks Teach Machines about Function Word Comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)* (pp. 235-249).

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. Annual Review of Linguistics, 7, 195-212.

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4902-4912).