Generating grammatically correct but improbable sentences (in Basque) / Gramatikoki zuzenak diren baina inprobableak diren esaldiak sortuz (euskararako)

Proposers/Proposatzaileak: Begoña Altuna and Itziar Gonzalez-Dios
Contact/Kontaktua: begona.altuna@ehu.eus; itziar.gonzalezd@ehu.eus

Description/Deskribapena:
Language models trained in larger datasets have proven to be very powerful in identifying language patterns and reproducing them in different scenarios. It has been argued that they learn word collocations as some words tend to appear together and it has also been argued that they learn how to do syntactic generalisations (Linzen and Baroni, 2021).
As a consequence, when trying to assess the grammatical knowledge that can be expected from a language model, the selection of the words and their position in the sentence play a crucial role in achieving a satisfactory analysis or outcome. Thus, trying to eliminate that bias may help to grasp the real grammatical knowledge a language model can mimic.
For this, we plan to generate a dataset in which word semantics will not interfere with grammatical knowledge, by means of word substitutions and word order alterations, for example. We plan to take advantage of the test sentences for language model evaluation.

Goals/Helburuak:
Creating a dataset of grammatically correct but improbable sentences using NLP techniques for the assessment of the language comprehension of the language models.

Requirements/Betebeharrak:
- Good knowledge of the target language (Basque)
- Knowledge of (Basque) morphosyntax
- Basic knowledge on WordNet and similar databases.
- Basic programming skills (e.g. Python for NLP)
- Basic knowledge of language models and neural classifiers (e.g. BERT)

Framework/Esparrua:
We will use Basque WordNet (Pociello et al., 2011), e-ROLda (Estarrona et al., 2016) and other semantic resources to substitute words in sentences (in Basque) so as to create sentences that will be grammatically correct but very unlikely to be present in training data.

Tasks and plan/Atazak eta plana:
- Selection of the sentence structures that will be the target of the experiment.
- Definition of the word substitution strategies.
- Generation of the test dataset.
- Evaluation of the language model through the test dataset.

References/Erreferentziak:

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In Proceeding3. https://doi.org/10.1145/3442188.3445922

Estarrona, A., Aldezabal, I., Díaz de Ilarraza, A., & Aranzabe, M. J. (2016). A methodology for the semiautomatic annotation of epec-rolsem, a basque corpus labeled at predicate level following the propbank-verbnet model. Digital scholarship in the humanities, 31(3), 470-492.Oxford University Press (Online ISSN 2055-768X - Print ISSN 2055-7671) doi: 10.1093/llc/fqv001

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

Pociello, E., Agirre, E., & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. Language resources and evaluation, 45(2), 121-142.