Exploring polysemy in specific academic domains / Polisemiaren azterketa alor akademiko espezializatuetan

Proposers/Proposatzaileak: Izaskun Aldezabal
Contact/Kontaktua: izaskun.aldezabal@ehu.eus

Description/Deskribapena:

Polysemy is one of the most difficult aspects to solve in multilingual tasks such as manual or machine translation. Although machine translation based on large language models has substantially improved, it has great room for improvement especially in selecting the right lemma in especific contexts, even more so in languages with few data. Being polysemy a broad field, in this work we want to explore polysemy cases in one (or more) specific academic domain(s) where a single lemma in one language is related to different lemmas in another language. It is preferable that one of the language involved is Basque, together with Spanish (and optionally other languages).

Goals/Helburuak:

The goal is to gather interesting sentence examples and word lists to create datasets for both the improvement of current results in machine translation and offering proposals in grammar and style correctors.

Requirements/Betebeharrak:

- Good knowledge of the languages involved
- Good knowledge of Garaterm and TZOS
- Basic knowledge of language models, their evaluation and improvement methods

Framework/Esparrua:

For the detection of polysemy cases, we will use TZOS. We will analyze polysemy cases in different domains, create significant datasets, and then evaluate the capability of language models before and after including the datasets in the process.

Tasks and plan/Atazak eta plana:

- Exploring polysemy cases in TZOS
- Building datasets.
- Evaluating and improving language models in machine translation
- Other possible applications

References/Erreferentziak:

María Jesús Aranzabe, Igone Zabala, Izaskun Aldezabal (2023). Goi-mailako testu akademikoak lantzeko baliabideak eta tresnak. II. CLARIAH-EUS workshop-a: Europako ikerketa azpiegiturekin lotuta egongo den euskararako ikerketa azpiegitura eraikitzen. Donostian, 2023ko azaroaren 23an.

Izaskun Aldezabal, Jose Mari Arriola, Arantxa Otegi (2022). TZOS: an Online Terminology Database Aimed at Working on Basque Academic Terminology Collaboratively. Proceedings of the 13th Language Resources and Evaluation Conference. Editors: Nicoletta Calzolari (Conference chair), Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Helene Mazo, Jan Odijk, Stelios Piperidis. http://www.lrec-conf.org/proceedings/lrec2022/index.html

Zabala, I., Lersundi, M., Leturia, I., Manterola, I., y Santander, G. (2013). GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea. En X. Alberdi, y P. Salaburu (Eds). *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira* (pp. 98-114). UPV/EHU argitalpen-zerbitzua.

Zabala, I., Aldezabal, I., Aranzabe, M.J., Arriola, J.M., Gonzalez-Dios, I, y Lersundi, M. (2018). Corpus-driven Terminology Work for Describing Basque Academic Terminology: the Weaving Terminology Networks programme (TSE programme). [Presentación de poster]. EAFT Terminology Summit, Donostia-San Sebastián, Spain.