Creation of a system for the extraction and normalisation of time expressions in Basque / Euskarazko denbora-adierazpenen erauzketa eta normalizaziorako sistema sortzea

Proposers/Proposatzaileak: Rodrigo Agerri and Begoña Altuna
Contact/Kontaktua: rodrigo.agerri@ehu.eus; begona.altuna@ehu.eus

Description/Deskribapena:
Time information helps to understand the information in the text, as it helps to place the events mentioned in the text in chronological order. Chronological points and durations are expressed in texts through time expressions. For instance, in Basque “2006an”, “gaur”, “5 egunez”, etc. In order to organise the temporal information present in text, time expressions need to be identified and a normalised value needs to be assigned to them.
Goals/Helburuak:
Adapting the TimeNorm tool for extracting and normalising time expressions in Basque.
Requirements/Betebeharrak:

- Good knowledge of Basque
- Knowledge of Basque morphosyntax
- Basic programming skills (e.g. Python for NLP)
- Basic knowledge of language models and neural classifiers (e.g. BERT)

Framework/Esparrua:
To detect time expressions, we will use neural networks and we will study zero-shot, few-shot, and supervised methods. For the normalisation we will evaluate rule-based and machine-learning methods and we will choose the most convenient.
Tasks and plan/Atazak eta plana:

- Development of a system for extracting time expressions in Basque.
- Development of a TimeNorm grammar to normalise time expressions in Basque.
- Evaluation of the system.

References/Erreferentziak:

Begoña Altuna, María Jesús Aranzabe, eta Arantza Díaz de Ilarraza. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. Linguamática, 6(2): 13–24, 2014a. ISSN 1647-0818. http://linguamatica.com/index.php/linguamatica/article/view/v6n2-1

Begoña Altuna, María Jesús Aranzabe, eta Arantza Díaz de Ilarraza. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0. Barne-txostena, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016, 2016a. https://addi.ehu.es/handle/10810/17305

Steven Bethard. A synchronous context free grammar for time normalization. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing vol. 2013 (2013): 821-826. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667674/pdf/nihms-619591.pdf

Nayla García. 2021. Multilingual Processing of Temporal Expressions: Detection and Normalization. Master Dissertation. University of the Basque Country.

Egoitz Laparra, Dongfang Xu, and Steven Bethard. 2018. From Characters to Time Intervals: New Paradigms for Evaluation and Neural Parsing of Time Normalizations. In: Transactions of the Association for Computational Linguistics 2018, Vol. 6, pp. 343–356 https://aclanthology.org/Q18-1025.pdf

Paramita Mirza and Anne-Lyse Minard. FBK-HLT-time : a complete Italian TemporalProcessing system for EVENTI-Evalita 2014. In: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014, pages 9-11. Pisa University Press, Pisa, 2014. http://digital.casalini.it/3044386

Dongfang Xu, Egoitz Laparra and Steven Bethard. 2019. Pre-trained Contextualized Character Embeddings Lead to Major Improvements in Time Normalization: a Detailed Analysis. In: Proceedings of The Eighth Joint Conference on Lexical and Computational Semantics. https://aclanthology.org/S19-1008.pdf