Computational Methods to Identify and Retrieve Passive and Relative Clauses: How to Apply it in Kurdish?

Hossein Hassani
University of Kurdistan Hewlêr
hosseinh@ukh.edu.krd

Relative and passive clauses have been studied from various computational perspectives. For example, relative clause simplification is an application that both Machine Translation (TM) and Text Summarization (TS) can use to provide a more smooth outcome. The information extraction on passive and relative clauses, using Information Retrieval (IR) tools, assists a variety of Natural Language Processing (NLP) tasks. That also enables linguists to study and compare the usage of the mentioned structures in a wide range of different texts, speeches, and contexts. For that, a large amount of effort is required. The task becomes more difficult for under-resourced languages. Kurdish, as a multi-dialect language, is considered a less-resourced language. To computationally process relative and passive clauses in Kurdish texts (and speeches), the researchers need different tools and data. The Kurdish-BLARK project (Hassani, 2018) aims to prepare the data and develop the tools that facilitate Kurdish processing. Since it was established, it has conducted several projects, and all outcomes have publicly been made available. However, as it has not been able to address all fundamental (and perhaps primitive) requirements yet, the tasks of passive and relative clauses have not been deemed to have high priority. ETPOLL-22 is an opportunity to discuss this track of research in Kurdish processing. Currently, several resources are available on the Kurdish-BLARK (e.g., a corpus on K-12 educational books, a corpus on Kurdish folklore, and parallel corpora among Kurdish (Sorani/Central Kurdish), Kurdish (Kurmanji/Northern Kurdish), and English) that could be used to study relative and passive clauses. Computational approaches can facilitate the comparison of frequency and the differences in the usage of those structures in various dialects of Kurdish. For example, a pattern matching approach along with a syntactic analysis (Yangarber et al., 2000; Yangarber and Grishman and Tapanainen, 2000; Zouaq and Gasevic and Hatala, 2012) or Machine Learning (Etzioni, 2011) could be adapted and applied on the available Kurdish corpora to study the efficiency of those approaches in the extraction of the mentioned structures.

References:

Etzioni, O., Fader, A., Christensen, J. and Soderland, S., (2011). June. Open information extraction: The second generation. In Twenty-Second International Joint Conference on Artificial Intelligence.

Hassani, H. (2018). BLARK for Multi-dialect Languages: Towards the Kurdish BLARK. Language Resources and Evaluation, 52(2):625–644.

Yangarber, R., Grishman, R., Tapanainen, P. and Huttunen, S. (2000). Automatic acquisition of domain knowledge for information extraction. In COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics.

Yangarber, R., Grishman, R. and Tapanainen, P., (2000). Unsupervised discovery of scenario-level patterns for information extraction. In Sixth Applied Natural Language Processing Conference (pp. 282-289).

Zouaq, A., Gasevic, D. and Hatala, M., (2012). Linguistic Patterns for Information Extraction in OntoCmaps. WOP, 929.