On the Current State of Kurdish Language Processing

Sina Ahmadi

Introduction

Natural language processing (NLP) and computational linguistics (CL) are of increasing importance in our information technology era. Given the current advances in applying artificial intelligence approaches for processing linguistic data, less-resourced and endangered languages can benefit to be documented and further processed by computer.

Despite its remarkable number of 20-30 million speakers, the Kurdish language has not received much attention in this realm. Recently, Ahmadi (2020) carried out an extensive study of the scientific contributions to the fields of NLP and CL for the Kurdish language demonstrating that progress in Kurdish language processing has been hindered by lack of open-source approaches inaddressing NLP tasks. Consequently, Kurdish language is still at its first stages of development inlanguage technology. In addition, to remedy this, the paper presents the Kurdish Language Processing Toolkit to address basic language processing tasks, particularly, text preprocessing, tokenization, transliteration, spelling error detection and correction and, morphological analysis.

Objectives

The current position paper aims at providing a roadmap for the future of language technology forthe Kurdish language by considering the progress that has been made so far. To this end, we include the Sorani and Kurmanji dialects, as our target dialects of Kurdish, and propose medium-term objectives in three categories of tools, resources and applications. We believe that addressingthese goals is essential to achieve more substantial progress and ultimately, can pave the way for the Southern Kurdish dialect and Zaza-Gorani languages to be also processed computationally.

A Proposed Roadmap

Considering the strategies for promoting NLP and CL for less-resourced languages described in (Allah & Boulaknadel, 2012), we propose a roadmap, illustrated in Figure 1, for the medium-term[1] objectives in Kurdish language processing. Moreover, the following criteria are taken into account:

(1) The inexistence of previous study of the kind for Kurdish language, based on the publicly available research and data.
(2) The priority of the tasks to be addressed in an NLP pipeline (Bird et al., 2009)
(3) The feasibility of the tasks given the current progress, particularly regarding annotated linguistic resources.

The components of this roadmap are mutually related in such a way that creating an application would require a tool depending on the existence of a resource. To the best of our knowledge, the described tasks have not been addressed for both Sorani and Kurmanji dialects. Therefore, they are proposed as future objectives for Kurdish language processing.

[1] As of April, 2021are proposed as future objectives for Kurdish language processing. Given the experimental natureof many NLP applications, the performance of contributions to the field would be subject of furtherstudies in the future.

Given the experimental nature of many NLP applications, the performance of contributions to the field would be subject of furtherstudies in the future.

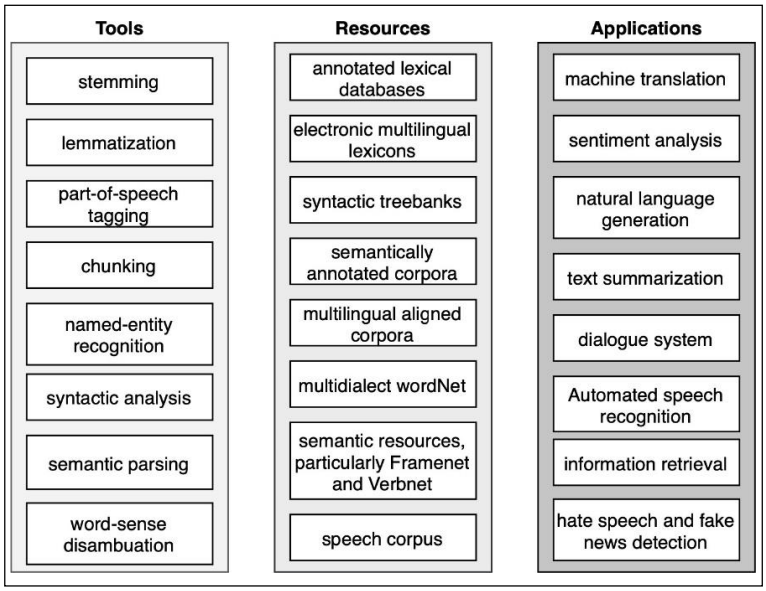

Figure 1. A proposed roadmap for medium-term progress in Kurdish language processing

Considering the hierarchical nature of many NLP tasks, as in machine translation which requires underlying tools such as tokenization and named-entity recognition, we believe that the proposed roadmap is beneficial to many other downstream applications, such as human-machine interaction.

Finally, it should be highlighted that annotation and creation of language resources is a costly andtime-consuming task. Therefore, we believe that Kurdish linguists and interested communities should play a more active role in integrating this endeavor in an openly accessible way.

Conclusion

This paper proposes a roadmap for future developments in Kurdish language processing by takingthe latest advances in the field into account. We believe that the proposed roadmap increases sustainability and effectiveness of future research in the Kurdish language processing and computational linguistic fields. Releasing projects under open-source licenses would accelerate future advances as well.

References

Allah, Fadoua Ataa, and Siham Boulaknadel. "Toward computational processing of less resourced languages: Primarily experiments for Moroccan Amazigh language." *Text Mining. Rijeka: InTech*(2012): 197-218.

Ahmadi, Sina. "KLPT–Kurdish Language Processing Toolkit." *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. 2020.

Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python:analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.