Neural Machine Translation for Arabic Language

Amal El Fallah Seghrouchni, Youssef Hmamouche

Key-words: Neural machine translation, NLP, Deep learning, Arabic language.

1 Research Motivation

Neural Machine Translation has become the most useful and the most powerful way to perform automatic translation. Nowadays, there are still some issues when translating dialects, in particular, Arabic dialects [1]. The Moroccan dialect is a specific language that is composed of several languages, and developing tools for translating the Moroccan dialect is crucial for the digitization process in Morocco.

2 Problem Statement

The Moroccan dialect lacks orthographic and grammatical syntax compared to a normal language, which makes it more difficult to learn. The lack of dictionaries and translation materials is also a factor that makes this language more difficult despite some efforts to gather data sets and materials related to Arabic dialects [2]. Nevertheless, it is quite close to the standard Arabic language, and the idea is to take advantage of that to help modern translation systems to understand partially the Moroccan dialect.

3 Research Scope

The first objective of this thesis is to develop an NLP system for Arabic language translation. Sequence-to-sequence (Sec2Sec) translation models based on recurrent neural networks (e.g., LSTM), and on self-attention (Transformers) will be discussed and investigated [3, 4]. Then, new architectures have to be proposed in order to improve state-of-the-art results. The second objective of the thesis is to propose an approach to translate partially standard Arabic to the Moroccan dialect, and eventually develop a *Chatbot* that includes this feature.

4 Admission Criteria

The PhD position is proposed by the International Center of Artificial Intelligence of Morocco, of the Mohammed VI Polytechnic University. Applicants with excellent cursus must be holders of a Master's, an engineering or an equivalent recognized degree in Computer Science or Applied Mathematics. In addition, they should have skills in Programming (Python and C++) and good communication

skills in English. Particular attention will be given to the suitability of this research project with the applicant's background.

References

[1] Salima Harrat, Karima Meftouh, and Kamel Smaili. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273, 2019.

[2] ElMehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. An open access nlp dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*, 2021.

[3] Ashmari Pramodya, Randil Pushpananda, and Ruvan Weerasinghe. a comparison of transformer, recurrent neural networks and smt in tamil to sinhala mt. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 155–160. IEEE, 2020.

[4] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.