УДК 004.8

METHODS AND ALGORITHMS FOR INTELLECTUAL ANALYSIS OF MEDICAL TEXTS FOR HEALTHCARE IN SYRIA

Jaafar Hammoud (ITMO University)
Scientific adviser - candidate of technical sciences, senior researcher Natalia Gusarova (ITMO University)

Introduction. Despite the spread of natural language processing (NLP) methods and their applications to the Arabic language, medical texts did not receive the same interest as general texts such as political, sports and economic news. This is due to the lack of sufficient medical datasets, and immaturity of the methods used with the Arabic language. The Arabic language in terms of processing is much more complex than the rest of the languages for many reasons, including:

1. It has three different forms (Classical, Modern Standard Arabic, and Dialectal)
2. There are about 12 million Arabic words in compare up to 600k words in English.
3. Letters in Arabic language have different shapes according to different positions in the word.
4. The rich and complex grammatical and spelling structure (Orthographic Ambiguity, Morphological Richness, Dialectal Variation, Orthographic Inconsistency).
5. Pronouns in Arabic are written in connection with the verb, and others are not apparent in writing and are clear through the meaning and context.

In this abstract, we present the main results that we have reached as a result of our work on Arabic medical texts and our use of modern techniques in machine learning and deep learning, such as Transformers and graph neural networks.

Main part. To overcome the first and main obstacle, which is the lack of datasets, we collected Arabic medical dataset with a team of volunteer medical students and doctors working in Syria, and then we applied traditional machine learning methods, deep learning methods, and graph neural networks. We can summarize the work with the following main points:

1. We have collected an Arabic medical dataset for the task of classifying Arabic medical texts. The dataset consists of two thousand documents in addition to the existence of ten medical categories (Cardiovascular, Blood, Liver, Nephrological, Eye, Bone, Ear, Gastrointestinal, Endocrine, and Immune) diseases. [1,2]

2. We applied the latest deep learning methods with texts, which are Transformers such as (Bert, AraBert, ABioNer, and GPT) networks, where we fine-tuned them on our dataset, and we got very good results compared to the same task for other languages. [2]

3. We introduced a new optimization method based on conjugate gradient methods. We used it instead of the Adam algorithm in the last layers that we adjusted, and we got the same accuracy, but with a time of 18% less. [3]

4. We applied graph neural networks. Despite their power and ability to represent complex structures, we have not seen any previous use of them with Arabic in the available literature. We took advantage of the grammatical richness of the Arabic language. We used additional documents that contain the syntax and description of the grammatical documents and tokens that we have in the dataset then embedded them together during the training of the network to obtain the highest accuracy in this task with the Arabic language.

Conclusion. The use of deep learning methods gives very satisfactory results with Arabic medical texts, but the Arabic language and because of its grammatical richness, the use and inclusion of its grammar in some way within the network training lead to better results. This is evident using Arabic grammar with graphic neural networks known for their ability to represent complex structures of knowledge.

References

1. Hammoud, Jaafar, Natalia Dobrenko, and Natalia Gusarova. "Named entity recognition and information extraction for Arabic medical text." Multi Conference on Computer Science and Information Systems, MCCSIS. 2020.

2. Hammoud, Jaafar, et al. "New Arabic medical dataset for diseases classification." Intelligent Data Engineering and Automated Learning–IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings 22. Springer International Publishing, 2021.

3. Hammoud, Jaafar, et al. "Using a new nonlinear gradient method for solving large scale convex optimization problems with an application on Arabic medical text." arXiv preprint arXiv:2106.04383 (2021).