Deep learning for medical information extraction - IXA group

Proposers: Arantza Casillas, Maite Oronoz, Alicia Pérez

Research group: IXA (http://ixa.eus)

Contact: arantza.casillas@ehu.eus

Description

This is a **research** project devoted to Medical Decision Support Systems (e.g. a commercial counterpart of such a system is IBM Watson) with expertise in **clinical text mining**, a sub-field of Artificial Intelligence applied to language understanding.

Clinical information extraction poses several **challenges** for algorithms based on inference from data: on the one hand, the system should be accurate; on the other hand, specific clinical data sources tend to be scarce and, thus, learning algorithms should be robust and be able to leverage inference in extremely adverse situations.

The aim of this project is to cope with situations such as learning about rare diseases. Data Science poses statistical tools and metrics to measure relatedness within numerical data. The key issue is, thus, to represent language in such a numerical space. To this end, deep learning emerged as a way to detect inherent features from language and represent texts in an abstract inferred space. Moreover, this scenario enables to link information in different languages, leveraging, thus, inference problems with few data from mono-lingual to multi-lingual knowledge representation. This framework accounting for statistical similarity enables us to find information about events (e.g. rare diseases) in several languages. There are closely related applications of similar techniques able to detect plagiarism seen as relatedness between text-styles. Hence, even though the project is applied to the medicine, its application scope is beyond it.

Learning outcomes: the student will acquire expertise in clinical text mining in the following areas:

- deep learning applied to textual inference for complex representation extraction
- relatedness and confidence metrics for artificial intelligence in text understanding

This project is defined in the context of the 3-year ProsaMed project http://ixa2.si.ehu.eus/prosamed/ involving 3 universities.

Goals

The student will apply deep learning techniques and relatedness metrics in order to build a prototype able to identify related documents. The key objectives are:

1) Analysis of the state of the art techniques for identifying similar elements in comparable data.

2) Design of a prototype for linking related documents written either in the same or in different language.

3) Implementation and evaluation of the prototype.

Requirements

- Good programming skills.
- It is recommended to take the course "Seminar on language technologies. Deep Learning"
- The master dissertation can be written in English, Spanish or Basque.

Framework

Python

Tasks and plan

Dec-Janunary: Study literature, get familiar with Python specific libraries and make use of relatedness metrics; from the very beginning, start to write down the thesis (antecedents, challenges, goals, data analysis).

February: Attend course "Seminar on language technologies. Deep Learning", familiarise with Pytorch. Development of deep representations of documents.

Mar-May: Development of language unaware textual representations and experimental layout; keep writting the thesis (experimental framework).

June: Write down the thesis (results and conclusions) and prepare the presentation

References

Laburu, M., Pérez, A., Casillas, A., Goenaga, I., & Oronoz, M. (2018, December). Can I find information about rare diseases in some other language?. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2102-2108). IEEE.

Mikel Artetxe, Gorka Labaka, Eneko Agirre (2018) Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) pages 5012-5019.

Pytorch: https://pytorch.org/

Research Software Platform. arXiv preprint arXiv:1705.06476. Retrieved from http://arxiv.org/abs/1705.06476

RECOMMENDED COURSE: Seminar on Language Technologies. Deep learning.

Deep Learning neural network models have been successfully applied to natural language processing. These models are able to infer a continuous representation for words and sentences, instead of using hand-engineered features as in other machine learning approaches. The seminar will introduce the main deep learning models used in natural language processing, allowing the students to gain hands-on understanding and implementation of them in Tensorflow.

Topics

- Introduction to machine learning and NLP with Tensorflow Deep learning
- Word embeddings
- Language modeling and recurrent neural networks
- Convolutional neural networks
- Attention mechanisms

Prerequisite. Basic programming experience, a university-level course in computer science and experience in Python. Basic math skills (algebra or pre-calculus) are also needed.