Overview on Information Extraction

Abilev Nurmukhamed
Kazakh-British Technical University
Faculty of Information
Technology
Almaty, Kazakhstan

I. First paper

The first paper presents its main idea as using Lisp architecture for LAPNLP, the portable NLP systems that integrates multiple customized, in-house developed and standard NLP tools, for clinical notes processing. Their system promotes easy data portability in multiple data systems by using boosted CDM – Common Data Model. So they developed LAPNLP phenotyping system which stores annotations from NLP tools into stand-off annotation format to standardize main data elements and ensure portability. The authors mentioned the advantage of LAPNLP over file system based pipelines such as BioC and UIMA CAS, and OMOP CDM. LAPNAP also built an interval tree to support and provide fast and efficient queries. LAPNAP workflow can be divided into several key points: 1)finding structures in documents, 2)breaking documents into sentences, sentences into tokens, 3)tagging them with part-of-speech, 4)grouping or chunking tokens, 5)deep parsing, 6)identifying UMLS annotations on tokens and phrases, 7)recognizing assertions as negations, 8)generating features and various machine learning components.

II. Second paper

The second paper represents its idea as study power Chinese text information extraction method from power fault countermeasure text. Author mentioned that electric power big data has main problem as diversity, non-structured data in power system. They want to find the way to improve fault handling methods by information extraction from fault countermeasure text by using NLP, without manually extraction of critical information of corresponding failure which has several problems. The text pre-processing has several steps: 1)Ontology Lexicon establishment, 2)Punctuation based sentence segmentation. Then text extraction goes next: 1)Text is cleaned to normalize structure, 2)Long text is divided by comma positions, to separate parsing phrases, 3)The main information extracted by matching with information extraction rule template of fault preplan text. They constructed the stop word library and power dispatching dictionary to guide the process by NLP technologies. Secondly they divided long text into short text and parsed them separately, semantic analysis chooses the master rules for extraction template. At the end, the key information about critical fault from countermeasure text are compared to rule template, and matched information can be extracted.

III. Third paper

The third paper has its main idea to overview an Biomedical information extraction methods and using NLP. The author says that in Biomedical information extraction or BioIE is a crossroad of Natural Language Processing, Medicine and Biology. BioIE has different tasks which requires using of new NLP technologies such as NER and Relation Extraction. Author emphasizes three main problems in BioIE and says that they are similar to problems in information extraction itself. They are: 1)Named entity recognition, 2)Relation extraction, 3)Event extraction. 1.Named Entity Recognition or NER is the biomedical part which includes study and recognition of proteins, diseases, treatments, drugs and e.t.c which grouped as entities. In this case information extracted from corpus by ontology, and has characteristic challenges: Synonomy, Abbreviations, Entity names has many variants.
2.Relation extraction involves the finding of related entities and they have many difference between each other such as protein-protein, treatment-treatment and e.t.c. Because of the tons of available medical information, it is impossible to extract relevant relations from published material for one person. It requires of creation related datasets of entities and using of NLP and machine learning techniques such as SVN, LSTM and CNN. 3.Event extraction of biomedical domain, which gained importance recently. Author says it is beyond version of Relation extraction. And become popular with availability of an annotated corpus with the “BioNLP’09. Shared Task on Event Extraction”. It involves prediction of trigger words over event types. It depends on dependency parsing and consists of three steps: 1)trigger detection, 2)argument detection, 3)semantic post. Compared to them, author described another work. There is “The Markov Logic Network” which predicts whether token is a trigger word, and the class its belong. Author mentioned that event extraction problems using MLN approach becomes feasible and linear in the sentence length. And fast progress in developing of BioIE systems will help researchers in biomedical domain.

References

[1] Luo, Y., Szlovits, P. (2018). Implemention a Portable Clinical NLP System with a Common Data Model – a Lisp Perspective.

[2] Sun, S., Dai, Z., Xi, X., Shan, X., Wang, B. (2019). Power Fault Preplan Text Information Extraction Based on NLP.

[3] Nair, S. (2017). A Biomedical Information Extraction Primer for NLP Researchers.