A SURVEY OF RELATION EXTRACTION TECHNIQUES USING HYBRID CLASSICAL AND STATE OF THE ART METHODS

A PREPRINT

Onur Kara
Hindsight Technology Solutions
South Plainfield, N.J.
okara83@gmail.com

Fei Chen
University of Minnesota - Twin Cities
Minneapolis, MN
feichen.yw@gmail.com

Tommy Jenkins
Sommd Entertainment Group
New York, NY
tommymjenkins@gmail.com

Daniel Hug
Hindsight Technology Solutions
South Plainfield, N.J.
danielpatrickhug@gmail.com

Sajedah Safari
Sommd Entertainment Group
New York, NY
ssafari@live.com

Seth Berger
George Washington University
Washington DC
seth.berger@gmail.com

ABSTRACT

The extraction of named entities and their relationships from unstructured biomedical texts has long been a topic of great interest. The challenges that arise when undertaking a project in the subdomain of biomedical text information extraction (IE) are well known, yet the influx of research projects and interest in the domain is accelerating rapidly. Much of this interest relates to recent advances in Natural Language Processing (NLP) methodologies (e.g., transformer and attention-based mechanisms). These remarkable results in IE often relied on large biomedical corpora (as with Bert, Biobert, Scibert, Glue, Biomegatron). In the present study, we prioritize time and cost-effectiveness by combining pipelined a series of classical and state-of-the-art NLP techniques, upstream and downstream, to various aspects of the greater biomedical relation extraction task. In addition to custom components, we primarily focused on the well-known enhancements implemented through open-source libraries. We obtained our first Named Entity Recognition (NER) result using the out-of-the-box transformer model provided from the popular NLP library, spaCy. We fine-tuned both Bert-base (3.3 Billion tokens, 1 million epochs)[1] and BioBertv1.2 (Bert-base corpus + 18 billion PubMed tokens 1 million epochs). Our goal was to establish reference scores for increasingly complex and expensive models. Specifically, by first using Bert-base, a general-domain model, we aimed to quantify the effects of utilizing transfer learning via the fine-tuning of the annotated domain-specific data. The study utilized multiple pre-and post-processing steps and varied combinations of such methods at different stages of the relation extraction pipeline. This survey examines the upstream effects of both classical techniques such as rule-based entity disambiguation and entity linking in order to normalize entities and relations. It also trains modern neural networks to perform entity disambiguation and coreference resolution for similar cross-referencing purposes and normalization. All pre-processing modules are examined in isolation as well as in combination with one another. We followed the performance of downstream tasks such as NER with RE and quantified them. In addition, to obtain meaningful and usable data and to counteract the many instances of sparsity encountered throughout the individual classes of relations initially provided, we performed Data Augmentation techniques via fine-tuning Generative Pre-trained Transformers (GPT-2)[4, 5]. The purpose of this data augmentation approach was to generate additional data points, i.e., sentences containing entity-relation-entity triples. We present this study as a survey that examines both classical and state-of-the-art methods for extracting drug-gene and gene-product relationships from unstructured biomedical literature. The main goal was to aid our understanding and the accessibility of the task of relation extraction and to determine which combination of techniques may provide the most promising path toward developing more sophisticated and cost-efficient hybrid architectures[2, 3].

Table 1. Language models pre-training information.

Model	Tokenizer	Vocabulary	Corpus	Domains	steps/epochs
BERT	WordPIece	30K	BookCorpus (2.5B tokens) + Wikipedia (0.8B tokens)	General	1M steps
BioBERT 1.1	WordPiece	BERT	BERT corpus + PubMed abstracts (4.5B tokens)	General + Biomedic	1M steps
BioBERT 1.0	WordPiece	BERT	BERT Corpus + PubMed abstracts (4.5B tokens) + PMC full-text articles (13.5M tokens)	General + Biomedic	470K steps
SciBERT	SentencePiece	30K	Semantic Scholar (3.17B tokens) (1.14M full text papers)	18% Computer Science and 82% Biomedical	Not reported
GPT-2	Byte Pair Encoding (BPE)	50k	8 million web pages, except Wikipedia (40 GB of text)	General	Not reported

References

[1] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL-HLT* (2019), pp. 4171–4186.

[2] Lei Huang et al. "EGFI: Drug-Drug Interaction Extraction and Generation with Fusion of Enriched Entity and Sentence Information". In: *arXiv preprint arXiv:2101.09914* (2021).

[3] Zhengbao Jiang et al. "CoRI: Collective Relation Integration with Data Augmentation for Open Information Extraction". In: *arXiv preprint arXiv:2106.00793* (2021).

[4] Yannis Papanikolaou and Andrea Pierleoni. "Dare: Data augmented relation extraction with gpt-2". In: *arXiv preprint arXiv:2004.13845* (2020).

[5] Jason Wei and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". In: *arXiv preprint arXiv:1901.11196* (2019).