Extracting Biomedical Relations from Text

Abiola Obamuyide and Blair Johnston
CMAC, University of Strathclyde, Glasgow
{first}.{last}@strath.ac.uk

Background

- The task is that of identifying entities (e.g. drugs, proteins, genes, diseases) and identifying relationships between them from e.g. scientific literature, clinical trial reports, etc.

- Useful for constructing pharmaceutical knowledge graphs, medicine repurposing/re-use, adverse medicine reaction detection, discovery of new medicines etc.

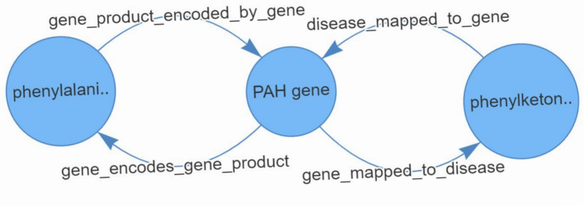

Figure 1: Knowledge Graph extracted from the following text:
"Aberrant splicing of phenylalanine hydroxylase mRNA: the major cause for phenylketonuria in parts of southern Europe. We report a mutation within the phenylalanine hydroxylase (PAH) gene that causes aberrant splicing of the mRNA and that is in tight association with chromosomal haplotypes 6, 10, and 36. Because of the high frequency of these particular haplotypes in Bulgaria, Italy, and Turkey, it appears to be one of the more frequent defects in the PAH gene causing classical phenylketonuria in this part of Europe. The mutation is a G to A transition at position 546 in intron 10 of the PAH gene, 11 bp upstream from the intron 10/exon 11 boundary. It activates a cryptic splice site and results in an in-frame insertion of 9 nucleotides between exon 10 and exon 11 of the processed mRNA. Normal amounts of liver PAH protein are present in homozygous patients, but no catalytic activity can be detected. This loss of enzyme activity is probably caused by conformational changes resulting from the insertion of three additional amino acids (Gly-Leu-Gln) between the normal sequences encoded by exon 10 and exon 11"

Data

- Existing datasets have limitations, e.g. assume a classification setting, are noisy, do not have annotations for end-to-end RE, etc.

- We introduce a new dataset suitable for end-to-end generative biomedical RE obtained from UMLS and Wikipedia.

UK Research and Innovation

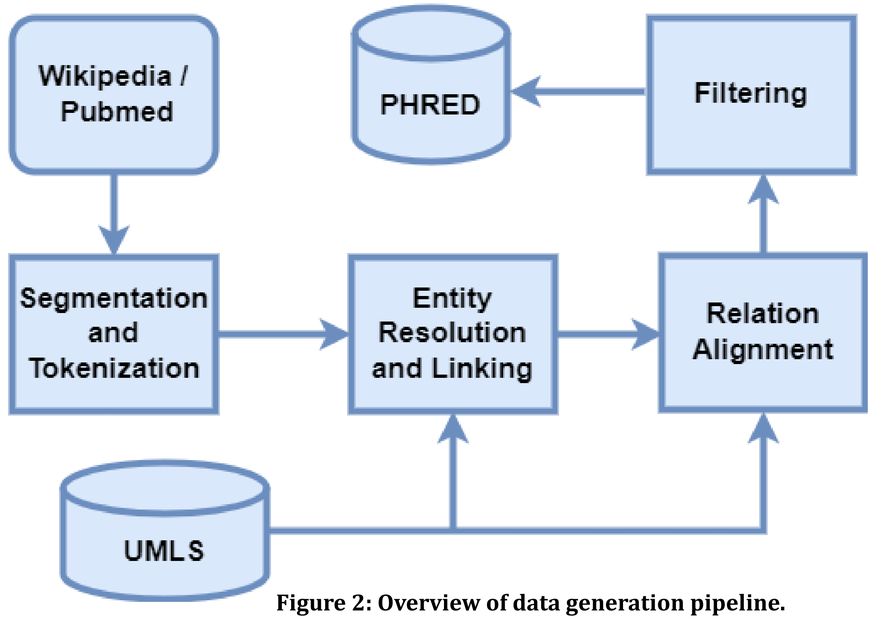

Figure 2: Overview of data generation pipeline.

Experimental Details

- Each instance in dataset consists of text (e.g. a sentence) together with all relation triples expressed in the text.

- Dataset has a total of about 107k instances which we split into 106k/500/500 train/val/test split. Baseline models include *BART*[1], *GENIE*[2], and *BIOGPT*[3].

- Our proposed approach combines elements from previous methods for true end-to-end generative pharmaceutical relation extraction.

Results

MODEL	Precision	Recall	F1
BART	23.16	14.95	18.17
GENIE	34.10	42.57	37.86
BIOGPT	39.02	41.10	40.04
OURS	**44.63**	**45.15**	**44.89**

Table 1: Results showing performance of our approach compared to other methods from the literature.

References

1. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*.
2. Josifoski, M., De Cao, N., Peyrard, M., Petroni, F., & West, R. (2022). GenIE: Generative Information Extraction. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference.*
3. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics.*
4. Cabot, P. L. H., & Navigli, R. (2021). REBEL: Relation Extraction by End-to-end Language generation. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021.*
5. Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Simperl, E., & Laforest, F. (2019). T-Rex: A large scale alignment of natural language with knowledge base triples. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 3448–3452.

DM² website

Acknowledgment: The authors would like to thank the Digital Medicines Manufacturing (DM²) Research Centre (Grant Ref: EP/V062077/1) for funding this work. DM² is co-funded by the Made Smarter Innovation challenge at UK Research and Innovation and partner organisations from the medicines manufacturing sector. For more information, visit cmac.ac.uk/dm2-home