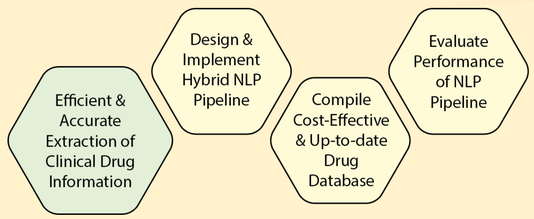
Combining Rule-Based Techniques and GPT-4 for Clinical Drug Information Extraction from SmPC Documents: A Natural Language Processing Approach to Developing Accessible and Up-to-Date Drug Databases

Malik Ahmed
Doses AI Limited, OpenPIL CIC

INTRODUCTION

The availability of accurate and up-to-date drug information is crucial for healthcare professionals, researchers, and patients. Summary of Product Characteristics (SmPC) documents contain essential information on drug properties, therapeutic indications, contraindications, and dosage guidelines. However, current drug databases can be expensive[2] and may not be updated frequently enough, leading to information gaps. Our research aims to develop a comprehensive approach that combines rule-based techniques and GPT-4 to accurately extract clinical drug information from SmPC documents and facilitate the creation of cost-effective, up-to-date drug databases.

OBJECTIVES



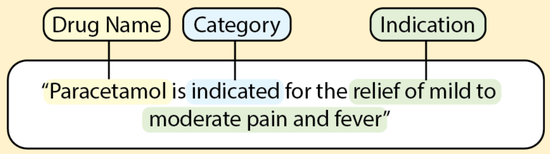
- Design and implement a hybrid NLP pipeline that integrates rule-based techniques and GPT-4 for extracting critical drug information from SmPC documents.
- Develop a cost-effective and up-to-date drug database by compiling the extracted data.
- Assess the scalability, efficiency, and applicability of the approach to other healthcare data sources and systems.

HYBRID-NLP PIPELINE

1. Data Collection & Pre-processing

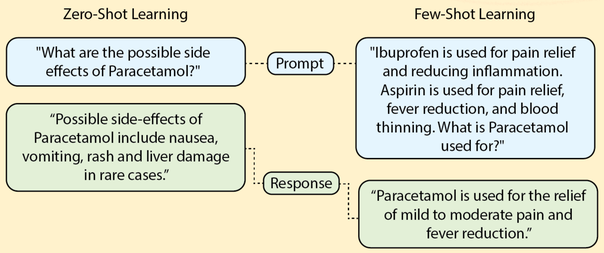
- Acquire SmPC documents from the European Medicines Agency (EMA) and other copyright-free public sources.
- Perform advanced pre-processing, including tokenisation and lemmatisation, specifically designed for pharmaceutical text data.

2. Rule-Based Information Extraction



- Implement token-level and sentence-level pattern matching using regular expressions and dependency parsing tailored to the pharmaceutical context.
- Design a comprehensive set of patterns to capture drug properties, therapeutic indications, contraindications, drug-drug interactions, and dosage guidelines.

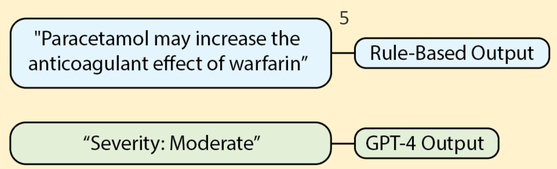
3. GPT-4 for Unsupervised Information Extraction



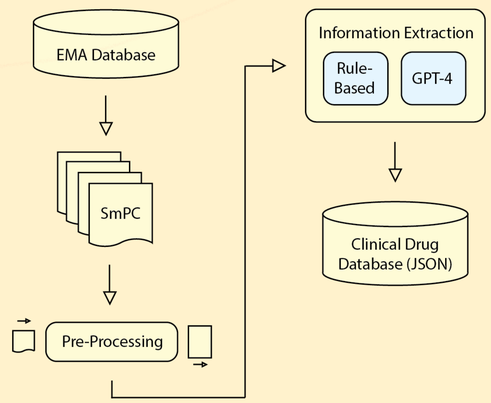
HYBRID-NLP PIPELINE CONT.

- Employ GPT-4 (LLM) to identify and extract relevant information from SmPC documents, utilising zero-shot or few-shot learning techniques[4].
- Implement active learning to iteratively improve the extraction performance with minimal supervision.

4. Drug Interaction Analysis



- Utilise GPT-4 to analyse the natural language literature within the SmPC document, identifying the severity of potential drug interactions.
- Apply rule-based techniques[3] to validate the accuracy of the detected interacting drugs by matching them with a predefined list of drug names and synonyms.



ROADMAP

- Optimise the hybrid NLP pipeline based on preliminary results and feedback from domain experts.
- Expand the approach to incorporate additional data sources, such as scientific publications, clinical trial records, and regulatory documents.
- Develop a user-friendly interface to query and visualise the compiled drug database.
- Integrate the drug database with existing electronic health records (EHR) systems and other healthcare applications.
- Assess the potential for extending the approach to additional languages and regions, enhancing its global applicability.

REFERENCES

1. Rubrichi, S. and Quaglini, S. (2012). Summary of Product Characteristics content extraction for a safe drugs usage. Journal of Biomedical Informatics, 45(2), pp.231–239.
2. Schlander, M., Hernandez-Villafuerte, K., Cheng, C.-Y., Mestre-Ferrandiz, J. and Baumann, M. (2021). How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. PharmacoEconomics, 39(11). doi:https://doi.org/10.1007/s40273-021-01065-y.
3. Kang, N., Singh, B., Afzal, Z., van Mulligen, E.M. and Kors, J.A. (2013). Using rule-based natural language processing to improve disease normalization in biomedical text. Journal of the American Medical Informatics Association, 20(5), pp.876–881. doi:https://doi.org/10.1136/amiajnl-2012-001173.
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
5. Mahe, I., Bertrand, N., Drouet, L., Simoneau, G., Mazoyer, E., Bal dit Sollier, C., Caulin, C. and Bergmann, J.F. (2005). Paracetamol: a haemorrhagic risk factor in patients on warfarin. British Journal of Clinical Pharmacology, 59(3), pp.371–374. doi:https://doi.org/10.1111/j.1365-2125.2004.02199.x.

CONTACT

Name: Malik Ahmed
Affiliation: Founder of doses.ai and OpenPIL
Contact: malik@doses.ai