Leveraging Large Language Models to Extract Terminology

Julie Giguere Anna Iankovskaia

julie.giguere@andovar.com anna.iankovskaia@andovar.com

Andovar Pte

Abstract

Large Language Models (LLMs) have brought us efficient tools for various natural language processing (NLP) tasks. This paper explores the application of LLMs for extracting domain-specific terms from textual data. We will present the advantages and limitations of using LLMs for this task and will highlight the significant improvements they offer over traditional terminology extraction methods such as rule-based and statistical approaches.

1 Introduction

In the context of localisation projects, extraction of terminology is more often than not the first linguistic task we perform. This is especially true as the type of localisation projects handled by professional services tend to be higher in complexity and require more subject matter expertise (SME). The more general type of content is now handled with robust results directly with the use of neural network machine translation (MT) or LLMs. The importance of technical terminology extraction in various fields has long been established. Having a solid glossary of terms is helpful to have a basis to agree on with the relevant stakeholders ahead of doing the translation. It is a helpful tool for the linguists and it allows to control the quality of the localised document via the use of term base and Computer Assisted Tools (CAT). It contributes to information retrieval and knowledge management. In the context of other natural language processing applications, term extraction is useful for text summarisation, semantic understanding and other usage. This paper explores the potential of LLMs for terminology extraction, providing an analysis of their effectiveness in comparison to traditional methods.

2 Context

LLMs have emerged as powerful tools in part because of their ability to capture complex language patterns and context. The architecture and training process of LLMs such as GPT-4 by OpenAI, BERT, and others helped us identify their potential for extracting terms effectively. We considered different methods for terminology extraction using LLMs. Contextual word embeddings -using pre-trained LLMs to generate contextualised word embeddings to capture word semantics based on surrounding text and enable better term identification. Sequence Labeling tasks ran with LLMs such as Named Entity Recognition (NER) to identify and extract domain-specific terms. Domain-specific fine-tuning to adapt pre-trained models to domain-specific corpora to improve term extraction accuracy within specialised texts.

3 Goal

Traditionally, in the context of a Language Service Provider, the extraction of terms from textual data was done using a CAT feature. We use Phrase inbuilt term extractor in our regular workflow and for the purpose of this comparison. The list of terms is then cleaned-up manually by a linguist to eliminate words that are not useful and to then add some terminology present in the data but omitted by the automatic extraction. This step is lengthy and requires the skill of a specialised resource such as a terminologist or a linguist with the relevant

domain expertise. The main goal of this paper was to find out if the use of LLMs could lower the human effort needed in extracting terminology.

4 Hypothesis

Statistical tools are frequency-based. In this experiment, a word should appear either at least once or at least two times in the text in order to be extracted. With a minimal frequency of 1, this results in a big number of false candidates. Their clean-up might take significant time. On the other hand, the tool with a minimal frequency of two generates less noise but it also omits numerous valid candidates. It is expected that LLM will outperform the other two methods by generating the highest number of valid terms with a minimum of noise therefore reducing the human effort needed in the glossary creation workflow.

5 Methodology

For this paper we are comparing a LLM (GPT- 4 by OpenAI) with a statistical model (Phrase term extractor). Phrase was used in two modes:

- '411' – from 1 to 4 word-terms; minimal frequency – 1; shortest word – 1 character.
- '421' – same but minimal frequency – 2.

The terms were extracted with Phrase 411, Phrase 421 and GPT-4 from three text samples and covering three domains: legal, medical, technical. Then, we also performed a manual extraction of the terms for each text sample and domain using a human specialist. This human extraction serves as a reference to identify the correct list of terms that should feature in the glossary. For each term extraction method, the number of correctly extracted terms (defined as a term also present in the list created by the human specialist) was divided by the total number of candidates proposed by the tool. The tools were compared based on the obtained accuracy score.

Prompt for the LLM extraction: content = "Extract domain-specific terms (domain: " + domain+") from the provided "+language+" text. \nText: "+text+"\nPlease extract only terms that are directly related to "+domain + " AND are present in the text (do NOT make up new terms, only extract existing ones in the text). Keep the same grammatical form and capitalization as these terms have in the text. Do not enumerate them in the output.``

6 Results

LLM got the highest accuracy score among the three tools (see Table 1). It generated one false term for technical (the term was not an existing term) and omitted some valid terms. Nevertheless its 'valid candidates vs. noise' coefficient was the best. We have been able to test in live projects and measure the saving in terms of human effort to clean up and finalise the list of extracted terms produced by the LLM-based approach versus the list of terms produced by the statistical model that are normally embedded into CAT software. This saving is important in the context of an LSP as it allows to be more competitive and to save time on the project's timeline. Several case studies showcased the successful application of LLMs in extracting technical terminology.

Content Type	LLM	Statistical '411'	Statistical '421'	Manually extracted terms	Text volume
Legal	21 correct terms / 25 candidates = 0.84	33 terms / 1,421 candidates = 0.02	16 correct terms / 183 candidates = 0.09	33 terms	504 words
Medical	52 terms / 66 candidates =	62 terms / 1,701 candidates =	9 correct terms / 93 candidates =	64 terms	487 words

Technical	0.78 41 correct terms / 56 candidates = 0.73	0.04 60 terms / 2,205 candidates = 0.03	0.1 24 correct terms / 288 candidates = 0.08	60 terms	724 words

Table 1 - Results

7 Advantages

LLMs is more efficient at capturing complex language patterns, handling polysemy, and adapting to domain-specific corpora through fine-tuning. They produce a more accurate and comprehensive terminology extraction.

8 Challenges and Limitations

LLMs may encounter difficulties in handling out-of-vocabulary terms, domain changes within the dataset, and they need significant computational resources during fine-tuning.

Conclusion: The LLMs term extraction has a great potential to improve the workflows and reduce the human effort needed. More work needs to be done to explore domain-specific pre-training and more efficient fine-tuning strategies to mitigate LLM limitations in technical terminology extraction.

We encourage researchers and professional service providers to adopt LLM-based approaches for technical terminology extraction to enhance domain understanding. We demonstrated in our results that LLMs outperform traditional methods. LLMs terminology extraction is a robust NLP application.

References

Chalkidis, I., Fergadiotis, M., & Malakasiotis, P. (2020). Assessing BERT's Syntactic Abilities. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA.

Huang, H., Li, T., Liu, W., Wang, F., & Liu, H. (2021). Incorporating Domain-Specific Knowledge into Large Pretrained Language Models with Domain Adapter. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Online.

Kageura, K., & Marshman, E. (2019). Terminology extraction and management. The Routledge handbook of translation and technology, 1, 61-77.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2021). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In Bioinformatics, Volume 36, Issue 4, Pages 1234–1240.

Li, D., Xie, Q., Qiu, M., Xu, Q., & Wang, X. (2021). TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online.

Xu, J., Zhu, W., Liu, P., & Zhao, J. (2021). Structure-aware Neural Model for Entity Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online.