Long-tail Entity Analysis using LLama and SpEL

Yasmin Madani
madanish@ualberta.ca

Hadi Sheikhi
hsheikhi@ualberta.ca

Abstract

Identifying named entities in texts is essential for effective information retrieval. It becomes challenging when encountering less common entities that are absent from the existing knowledge base, known as long-tail entities. A previous study utilizes elementary tools for Named Entity Recognition and Entity Linking to outline a methodology for identifying long-tail entities. We follow the same approach, replacing their elementary tools with state-of-the-art models. We compare our model against that of the prior work, demonstrating that exploiting modern neural models enhances the performance of our methodology.

1 Introduction

Real-world objects are referred to as *entities*. News articles—typically formatted in natural language text—contain mentions of entities, categorized in different categories such as people, locations, and organizations.

Named Entity Recognition (NER), which identifies entity mentions in natural texts, has many important applications: (1) NER provides insights into document topics, particularly news categories. (2) NER plays a significant role in identifying semantically similar news articles (Stockem Novo and Gedikli, 2023) which can be employed directly in news recommendation systems and search engines. For instance, Gupta et al. (2021) verify news based on its similarity to other news. (3) Filtering relevant documents with respect to entities is an essential task in the context of knowledge base construction and maintenance (Reinanda et al., 2016). State-of-the-art document filtering techniques rely on the unique characteristics of individual entities to effectively distinguish them.

Long-tail entities, less common entities absent from the existing knowledge base, are particularly intriguing as they are less identified by Entity Linker (EL; Reinanda et al., 2016). To address this challenge, Esquivel et al. (2017) propose a methodology to detect long-tail entities in news articles by calculating the intersection of entities detected by statistical NER and EL. In particular, they employ Stanford NER (Finkel et al., 2005) and DBPedia Spotlight (Mendes et al., 2011) as NER and EL, respectively. Utilizing statistical models in the era of neural networks may not be the best practice.

We integrate the system proposed by Esquivel et al. (2017) with state-of-the-art neural network-based models for NER and EL. We demonstrate employing our approach significantly increases the overall performance of long-tail entity detection. Moreover, we conduct a comprehensive comparative analysis between our findings and those reported by (Esquivel et al., 2017).

2 Related Work

Named Entity Recognition Wang et al. (2023) proposes a novel approach to NER by introducing GPT-NER, which utilizes Large Language Models (LLM) to enhance the performance. This study bridges the gap between NER as a sequence labelling task and text-generation models. It is done by transforming NER into a text-generation task, which perfectly aligns with the capabilities of LLMs. The approach involves using special tokens to mark entities in the input text, thereby simplifying the extraction process. This method has shown significant performance improvement, particularly in low-resource and few-shot setups, suggesting its potential applicability to our work. Moreover, Wang et al. (2023) suggests a retrieval approach for a few-shot setting, prioritizing examples that are semantically similar to the input sentence.

Sentence transformers (Reimers and Gurevych, 2019) provides a unified framework that enables us to do semantic search among a corpus, by computing cosine similarity between sentence representations. Consequently, we employ the ConLL

2003 (Tjong Kim Sang and De Meulder, 2003) dataset as our example pool and perform a semantic search for each input sentence using sentence transformers (Reimers and Gurevych, 2019).

However, we believe a fine-tuned version of LLama2 (Touvron et al., 2023), with the same amount of parameters, would perform better than a pre-trained version. Zhou et al. (2024) illustrated their ability to faithfully replicate the capabilities of LLMs within a specific application class while maintaining their ability to generalize across various semantic types and domains. Employing NER as a case study, they effectively distilled these capabilities from LLMs into a significantly smaller model, called UniversalNER. It can recognize diverse types of entities and concepts in text corpora from a wide range of domains. Nonetheless, the utilization of such large models presents a significant challenge, notably the high generation load during inference, particularly when considering all entity types together. Lu et al. (2024) addressed this challenge by proposing parallel computations for named entities. Since entity types are independent of each other, they can be processed simultaneously without compromising performance (Lu et al., 2024).

Additionally, to mitigate the generation bottleneck, Lu et al. (2024) suggested that a long sequence (such as an article) can be partitioned into smaller sequences (i.e., sentences) without significantly impacting accuracy. Although we believe breaking articles into sentences is, actually, harming the performance, due to resource limitations, we choose to follow this prior work.

Entity Linking Neural network-based models have demonstrated exceptional performance in Entity Linking (EL), as well as various other natural language processing (NLP) tasks, due to their ability to learn useful distributed semantic representations of linguistic data (Young et al., 2017). For instance, Chen et al. (2020) proposed an entity linking (EL) model that jointly learns mention detection (MD) and entity disambiguation (ED). Their model utilizes task-specific heads on top of shared BERT contextualized embeddings. Mrini et al. (2022) defined entity linking as a sequence-to-sequence translation task, employing BART (Lewis et al., 2020).

Zhang et al. (2021) introduced entity linking as a question answering task. However, all of these methods are either resource-intensive or have a relatively long inference time. To address these challenges, Shavarani and Sarkar (2023) introduced Structured Prediction for Entity Linking (SpEL) wherein they could achieve the state-of-the-art on the commonly used AIDA (Hoffart et al., 2011) dataset. They used RoBERTa (Liu et al., 2019) as a base model for sub-word level prediction and extend these predictions to phrase-level predictions. The authors claimed that their proposed method is more efficient (it has fewer parameters) and accurate than the other methodologies in the literature. Consequently, we choose to utilize SpEL as our entity linker for this project.

3 Methods

3.1 Named Entity Recognition

Llama2-7b [Failed] In our pilot studies, we followed GPT-NER (Wang et al., 2023) using Llama2-7b[1] instead of GPT due to its open-source accessibility and cost-effectiveness. We provided few-shot demonstrations to Llama2 and prompt it to extract the desired entities in a given sentence. We conducted a semantic search within the ConLL 2003 dataset to improve our few-shot demonstrations by utilizing all-mpnet-base-v2[2], which is available in the sentence-transformers library (Reimers and Gurevych, 2019). This method resulted in issues. We realized that for some samples, Llama2 hallucinated, meaning that it considered non-desired entities as entity types. For instance, “Maccabi Tel Aviv” is a football team that has been marked as a location entity in the following example.

> Chelsea kicks off its Champions League campaign with an opening fixture against @@ **Maccabi Tel Aviv ##** on Wednesday evening

To address this issue, we initiated an experiment wherein the Llama2 model was prompted to perform self-verification (Wang et al., 2023). This approach mitigated the model’s hallucination. However, a new challenge emerged during our investigation. We observed that, in certain instances, rather than producing annotations as expected, the Llama2 model began generating input-output pairs. Unfortunately, we could not find an approach to tackle this issue, thus we chose to switch to the structured annotation output format. More details and prompts available in Appendix A.1.

[1] https://huggingface.co/meta-llama/Llama-2-7b

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

UniversalNER We utilize UniversalNER (Zhou et al., 2024) as our NER component and prompt it using a conversational approach, in which we provide the text and entity type during a conversation. In particular, we employ UniNER-7B-type. We expect the agent to output desired entities in a JSON format.

Post Processing UniversalNER considers pronouns and actors of a sentence as named entities. In our pilot studies, we realized that some common words (e.g., company, police, and banks) related to organizations have also been mentioned as named entities. A more detailed analysis of the frequency of lowercase words and pronouns (Figures 1, 2) shows that these cases are not long-tail entities and can be removed during the post processing. We remove pronouns in the Person category and lowercase entities in all 3 types if and only if all of their occurrences are lowercase. Note that all the lowercase detected entities can not be removed because some popular entities have occurred in their lowercase form as well (i.e., "twitter" instead of "Twitter").

3.2 Entity Linking

We utilize SpEL (Shavarani and Sarkar, 2023) for entity linking. We use their official repository[3] and the SpEL-base model. We utilized their checkpoint[4] where they consider 500k most frequent Wikipedia identifiers as well as the in-domain and out-of-domain data entity vocabularies to form the model output vocabulary.

4 Results

4.1 Data

Esquivel et al. (2017) used one million news articles sourced from news and blog sources in September 2015. Due to the time and computational resource limitations, we uniformly sampled 10% of data for our experiments. Respectively, we compare our results with 10% of statistics reported by Esquivel et al. (2017).

4.2 Evaluation Metric

Overlap Following Esquivel et al. (2017) we adopt a criterion for tag overlap, where two tags are overlapping if either the start or end offsets of one tag fall within the offsets of the other tag.

Parameter	Value
temperature	0
max_new_tokens	256
tensor_parallel_size	1

Table 1: Hyper-parameters for UniversalNER inference.

Method	PER	LOC	ORG
UniversalNER	**923512**	**605979**	**1152485**
Stanford NER	771000	552000	537000

Table 2: Number of detected entities. Stanford NER results are adopted from Esquivel et al. (2017). Since we used 10% of the dataset, we multiplied their results by 0.1. These results are reported on Person (PER), Location (LOC), and Organization (ORG) entity types.

4.3 Experimental Setup

We conduct all our experiments on Google Cloud Compute Engine, utilizing Tesla L4 GPUs with 23GiB memory, making them able to load large language models. For Large Language Model inference, we employed VLLM [5] library, an optimal library for LLM inference. Specific values of hyper-parameters is elaborated in Table 1.

For SpEL inference, we used the very exact hyper-parameters defined by the authors. We illustrate our experimental setup in Figure 3, Appendix.

4.4 Named Entity Recognition

We first compare the number of entities detected by UniNER (Zhou et al., 2024) against those detected by Stanford NER (Finkel et al., 2005). As shown in Table 2, our method detects more entities than its competitor across all three entity types. Especially, a significant improvement is demonstrated for the organizations. An example of entities detected by our method but missed by the Stanford NER is "Pandora Jewelry."

4.5 Entity-Linking

A comparison of the accuracy between our two entity linkers (SpEL and Spotlight) across the entire dataset is unattainable given the absence of ground truth for entity linking. Nevertheless, we conducted a manual evaluation of the entity linker results. We randomly selected 100 samples from the data and evaluated whether each entity linker correctly associates the named entity with the appropriate node in the graph. Notably, we define a failure for the linker only when it links an entity to

[3] https://github.com/shavarani/SpEL

[4] vault.sfu.ca/index.php/s/8nw5fFXdz2yBP5z/download

[5] https://github.com/vllm-project/vllm

Method	PER	LOC	ORG
UniNER+SpEL	**30.92**	59.09	34.53
Esquivel et al. (2017)	26.59	**64.55**	**39.44**

Table 3: Comparing same-type overlap percentage (%) over the chosen portion of data. These results are reported on Person (PER), Location (LOC), and Organization (ORG) entity types.

an incorrect graph node, whereas refraining from linking an entity is not considered a failure in terms of accuracy assessment. Based on the manual evaluation, we claim that SpEL is more accurate than Spotlight because SpEL can correctly link 78% of samples, while Spotlight can link only 59%. Our results show that even if Spotlight could link a high number of entities to the graph, the accuracy of these links is still questionable.

4.6 Overlap

Esquivel et al. (2017) report overlap between entity types in different settings. We focus on same-type overlap, where we just compute the overlap between entities of the same type. As illustrated in Table 3, our approach shows a higher overlap on Persons (PER); however on the two other entity types, Esquivel et al. (2017) shows a slightly higher overlap than ours. Considering the analysis in Section 4.5, we can claim that high overlap for Spotlight does not necessarily mean higher accuracy of linked entities. Therefore, while SpEL got lower overlap scores in some cases, based on the observations in Section 4.5, these linked entities are far more accurate than Spotlight outcomes.

5 Discussion

Our methodology shows relatively good performance on a special portion of data. Both our study and Esquivel et al. (2017) lack an automated measuring for the entity linker's accuracy. Moreover, we performed a case study for error analysis which is elaborated in the next section.

5.1 Error Analysis: A case study

To illustrate the egregious errors committed by Spotlight, consider the sentence "Ronaldo scored 8 goals for Brazil in World Cup 2002." Spotlight links the word "Ronaldo" to the famous football player, Cristiano Ronaldo, which completely mismatches the context. SpEL, however, links this word to "Ronaldo_(Brazilian_footballer)", which is a correct resolution for this word in the given context. This case study signifies the effectiveness of incorporating context for entity linking. Furthermore, it can be confirmed that Spotlight is prone to a higher incidence of irrelevant linking errors compared to SpEL when it links a greater number of entities. In other words, entities linked by SpEL are generally more reliable than those predicted by Spotlight.

6 Conclusion

In this study, we show that Llama 2 is not a capable method for the Named Entity Recognition task, even though we used the same method that worked on GPT (Wang et al., 2023). Moreover, there is no automated metric available to conduct the accuracy of entity linking on unstructured data, such as the Signal 1M dataset. Furthermore, utilizing a context-aware approach, such as SpEL, would perform significantly more accurately than a statistical model like Spotlight. Finally, we demonstrate that a distilled large language model for NER would perform outstandingly better than Stanford NER.

As discussed in Section 5, it is required to validate our methodology across the entire dataset. Moreover, providing a metric to compute the entity linker's accuracy is a requirement to conduct fair comparisons. Moreover, Large Language Models might be a potential candidate for the task of entity linking, removing the requirement for a knowledge graph for this task, given their broad general knowledge representation.

7 Repository URL

Our code is available Here.

8 Acknowledgement

We have used ChatGPT for grammatical checks and proofreading.

References

Haotian Chen, Xi Li, Andrej Zukov Gregoric, and Sahil Wadhwa. 2020. Contextualized end-to-end neural entity linking. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 637–642, Suzhou, China. Association for Computational Linguistics.

José Esquivel, Dyaa Albakour, Miguel Martínez, David Corney, and Samir Moussa. 2017. On the long-tail entities in news. pages 691–697.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Vishwani Gupta, Katharina Beckh, Sven Giesselbach, Dennis Wegener, and Tim Wirtz. 2021. Supporting verification of news articles with automated search for semantically similar articles. *ArXiv*, abs/2103.15581.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jinghui Lu, Ziwei Yang, Yanjie Wang, Xuejing Liu, Brian Mac Namee, and Can Huang. 2024. Padellmner: Parallel decoding in large language models for named entity recognition.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, page 1–8, New York, NY, USA. Association for Computing Machinery.

Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi, and Hamed Firooz. 2022. Detection, disambiguation, re-ranking: Autoregressive entity linking as a multi-task problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1972–1983, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2016. Document filtering for long-tail entities. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 771–780, New York, NY, USA. Association for Computing Machinery.

Hassan Shavarani and Anoop Sarkar. 2023. SpEL: Structured prediction for entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11123–11137, Singapore. Association for Computational Linguistics.

Anne Stockem Novo and Fatih Gedikli. 2023. Named entities as key features for detecting semantically similar news articles. *International Journal of Semantic Computing*, 17(04):633–649.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and E. Cambria. 2017. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.*, 13:55–75.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2021. Entqa: Entity linking as question answering. *CoRR*, abs/2110.02369.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition.

A Appendix

A.1 Llama2-7b

A.1.1 Llama2 Prompt

The prompt structure consists of three parts including 1) Task description 2) Few-Shot demonstrations 3) Input sentence.

According to (Wang et al., 2023) using special tokens @@## to annotate the desired output will reduce the generation load of the large language model because LLM can understand the use of these characters and copy the rest of the sentence. We prompted the model separately for each entity type because this makes it easier for the LLM to understand the task and it would perform more efficiently.

The following is the prompt for extracting Location-named entities. The same would be applied to the person and organization categories.

I am an excellent linguist. The task is to label Location entities in the given sentence. Below are some examples:
Input: Only France and Britain backed Fischler's proposal.
Output: Only @@France## and @@Britain## backed Fishler's proposal.
Input: Germany imported 47600 sheep from Britain last year.
Output: @@Germany## imported 47600 sheep from Britain last year.
Input: [SENTENCE]
Output:

A.1.2 Self-Verify Prompt

I am an excellent linguist. The task is to verify whether the word is a Location entity extracted from the given sentence.
The given sentence: Only France and Britain backed Fischler's proposal.
Is the word Britain in the given sentence a Location entity? Please answer with Yes or No.
Yes.
The given sentence: Germany imported 47600 sheep from Britain last year.
Is the word Germany in the given sentence a Location entity? Please answer with Yes or No.
Yes
The given sentence: [SENTENCE]
Is the word [WORD] in the given sentence a Location entity? Please answer with Yes or No.

A.2 UniversalNER

A.2.1 UniversalNER conversational prompt

Prompt: Text: [SENTENCE]
Agent: I've read this text.
Prompt: What describes [ENTITY TYPE] in the text?
Agent: ["NE1", "NE2",...]

A.3 Frequency Analysis

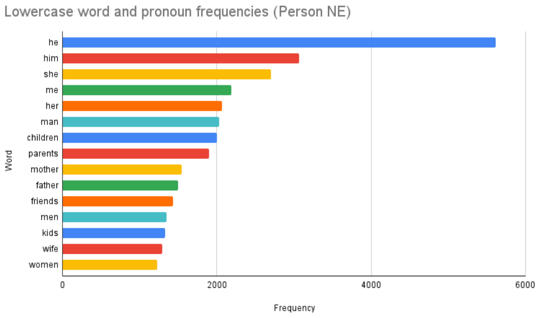

Figure 1: Lower Case And Pronoun Frequency For PERSON

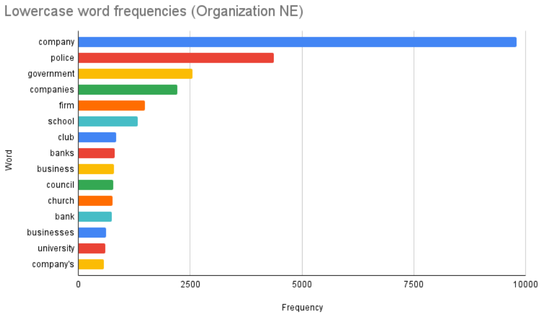

Figure 2: Lower Case Frequency For Organization

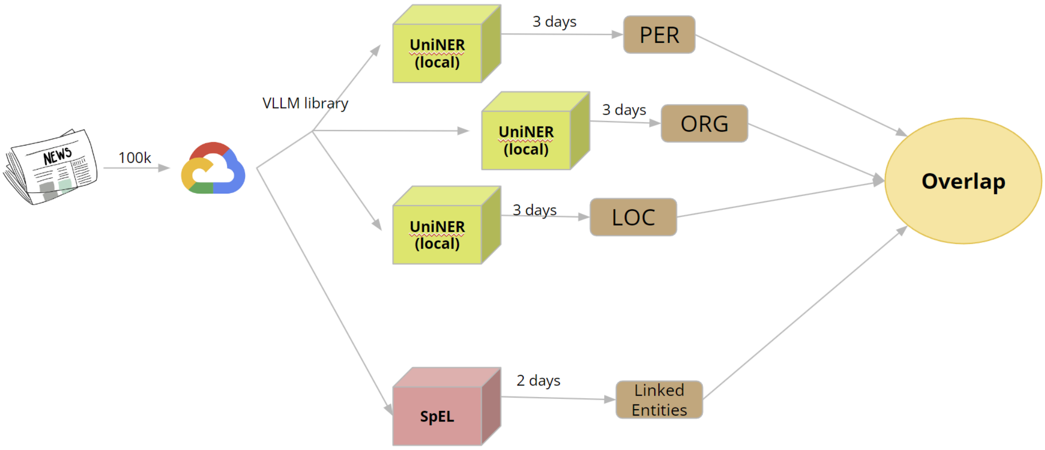

Figure 3: Experimental procedure of our methodology.