# YNU-HPCC at SemEval-2025 Task 2: Local Cache and Online Retrieval-Based method for Entity-Aware Machine Translation

**Hao Li, Jin Wang and Xuejie Zhang**
School of Information Science and Engineering,
Yunnan University
Kunming, China
haoli@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

## Abstract

This study reports the details of the Entity-Aware Machine Translation model proposed by the YNU-HPCC team to participate in the SemEval 2025, Task 2. The task focuses on Entity-Aware Machine Translation (EA-MT) to enhance the translation of sentences containing challenging named entities. We investigated traditional machine translation (MT) and large language model (LLM)-based approaches, evaluating their performance using metrics such as M-ETA, and COMET. We integrated a BERT-based Named Entity Recognition (NER) module for the traditional MT system. This approach is simple and intuitive, providing fast performance for common NE categories while achieving accuracy at the standard level. In the LLM-based system, we leveraged multiple LLMs and designed tailored prompts supplemented with a few examples, to guide the model in recognizing named entities. The system effectively translated these entities with high precision by incorporating contextual information from the rest of the sentence. A comparative evaluation of both methods aims to provide insights for future research. More details can be found here.[1]

## 1 Introduction

The definition of named entities was first proposed by Beth M. Sundheim in 1995 (Ide et al., 2003). Named entities typically refer to entities in a text that have a specific identity or name. In machine translation tasks, named entities may be related to language and cultural differences. The meaning of the same named entity can vary significantly in different contexts. The translation result may be unsatisfactory if the model fails to correctly identify potential named entities in a sentence and translate them in the context. In the past, people used methods such as regular matching, traditional machine learning models like SVM (Ju et al., 2011), and context-dependent models like LSTM for translating sentences with named entities. This paper proposes four methods, respectively, based on traditional machine translation models and LLM-based models, aiming to improve the translation accuracy of sentences with named entities and explore the capabilities of LLMs.

This study proposes a local cache and online retrieval-based method for translating sentences containing named entities, which consists of two modules. The NER module is primarily used to accurately identify named entities in the sentence, preparing for the subsequent translation task. The translation module is mainly responsible for integrating the NE translation results and translating the entire sentence in context. Experimental results show that traditional MT models have speed and deployment cost advantages, while LLM-based models outperform translation accuracy and contextual coherence.

The rest of the paper is organized as follows: First, our team propose four methods to improve the performance of ea-mt based on the two modules mentioned above. Then we test them across all ten languages. Finally, we will analyze the experimental results.

## 2 Related Work

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP), with applications in social media (Peng and Dredze, 2015), news (Vychegzhanin and Kotelnikov, 2019), e-commerce (Vychegzhanin and Kotelnikov, 2019), Nested Medical NER (Du et al., 2024) and other fields. NER in the past was performed using methods such as regular expression matching, traditional machine learning models like Support Vector Machines (SVM), Conditional Random Fields

[1] The code of this paper is available at: https://github.com/skyfuryonline/SemEval2025_task2_YNU-HPCC

*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 476–481
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

(CRF) (Panchendrarajan and Amaresan, 2018), Maximum Entropy models (MaxEnt) (Saha et al., 2008), Recurrent Neural Networks (RNN) (Lyu et al., 2017), and Long Short-Term Memory networks (LSTM). Although these methods were effective in specific domains, the training process was cumbersome, and the ability to recognize context remained weak. With the introduction of Transformers and the emergence of various pre-trained models (Liu et al., 2023), recent advancements in pre-trained models, such as transformer-based model in name entity recognition (Luo et al., 2022), BERT (Ma et al., 2021), Tacl-BERT, and Nezha (Wei et al., 2019), have significantly improved NER performance. In SemEval 2022 Task 11 (Chen et al., 2022), impressive results were achieved by two teams: one proposed a knowledge-based NER system, while the other enhanced their model's performance by utilizing a gazetteer constructed from Wikidata. In SemEval 2023 Task 2 (Lu et al., 2023), one of the top-performing teams approached NER as a sequence labeling problem. These methods have performed well in their respective tasks and provided valuable insights for our upcoming task.

## 3 System Description

This section introduces four methods implemented in Task 2 (Conia et al., 2025) and analyzes how each works.

### 3.1 Method 1: M2M-100 with BERT

The m2m-100 pre-trained translation model performs the overall translation task in this approach (Fan et al., 2021). In contrast, a BERT model pre-trained on the CoNLL-2003 Named Entity Recognition dataset is used for named entity recognition (Sang and De Meulder, 2003). A named entity dictionary is maintained to identify potential named entities in the prediction dataset using the pre-trained model, and their corresponding translations are retrieved from Wikidata.

Considering the model size, named entities in the sentence may not be correctly truncated. Therefore, the position information of each named entity is recorded, and truncation is extended to the nearest space, sentence start, or sentence end. Since potential parentheses, punctuation, numbers, and special symbols may interfere with identifying named entities, a set of regular expression rules is also defined to clean the entities. The recognized named entities' corresponding translation information is searched on Wikidata. Given the complex page information in Wikidata and the tendency for similar entries to point to the same meaning, a similarity algorithm is used for filtering. The top 20 entries are retrieved, and their similarity to the queried entity is calculated. The number of statements and site links are also considered to prevent interference from similar entries. The highest similarity entry translation is obtained by calculating a weighted overall similarity. Before being translated by the m2m-100 pre-trained model, the named entities are replaced using a fixed-length sliding window, and the result is input to the translation model to obtain the final translation.

### 3.2 Method 2: Qwen2.5-32B with M2M-100

Considering the impact of BERT's model size on named entity recognition (NER) accuracy, the NER module was replaced with the 32b Qwen model. Prompts were carefully designed to guide the model in identifying potential entities within the sentence and returning them as a list (or None if no entities were found). A set of examples, accompanied by step-by-step instructions, was provided to facilitate the model's understanding of the task. The resulting output is then directly provided as input to the m2m-100 model. Through these changes, the average M-ETA score for each language improved by 10 to 15 points.

However, since the translation task was based on the m2m-100 model, the improvement in COMET scores was limited. Furthermore, traditional MT models may alter or re-translate named entities to some extent, leading to potentially uncontrolled results. Fine-tuning a machine translation (MT) model on a given training dataset can help the model memorize translation rules for named entities to some extent. However, this approach has two potential drawbacks: first, the overall translation ability of the model may decrease, as indicated by a slight reduction in COMET scores for each language; second, the model is limited to memorizing the translation rules for named entities in the training dataset and lacks strong generalization ability, making it unable to handle unseen entities or new translation scenarios.

### 3.3 Method 3: Qwen Ner and Translator

Considering the factors above, the Qwen model replaced the m2m-100 translation model. The Qwen-Max API from the Qwen series expanded

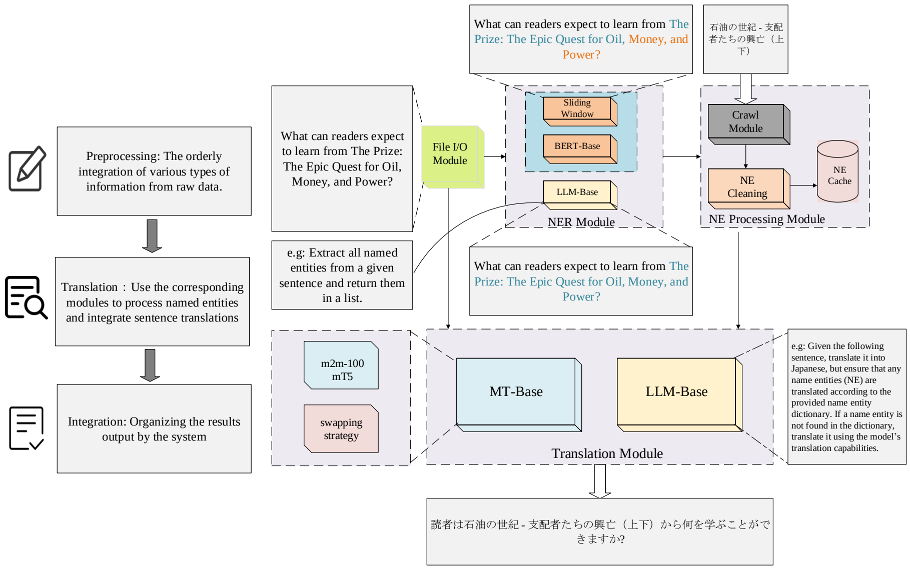

Preprocessing: The orderly integration of various types of information from raw data.
Translation：Use the corresponding modules to process named entities and integrate sentence translations
Integration: Organizing the results output by the system
What can readers expect to learn from The Prize: The Epic Quest for Oil, Money, and Power?
e.g: Extract all named entities from a given sentence and return them in a list.
File I/O Module
What can readers expect to learn from The Prize: The Epic Quest for Oil, Money, and Power?
Sliding Window
BERT-Base
LLM-Base
NER Module
What can readers expect to learn from The Prize: The Epic Quest for Oil, Money, and Power?
石油の世紀 - 支配者たちの興亡（上下）
Crawl Module
NE Cleaning
NE Cache
NE Processing Module
m2m-100 mT5
swapping strategy
MT-Base
LLM-Base
Translation Module
e.g: Given the following sentence, translate it into Japanese, but ensure that any name entities (NE) are translated according to the provided name entity dictionary. If a name entity is not found in the dictionary, translate it using the model's translation capabilities.
読者は石油の世紀 - 支配者たちの興亡（上下）から何を学ぶことができますか?

Figure 1: The overview of the system architecture

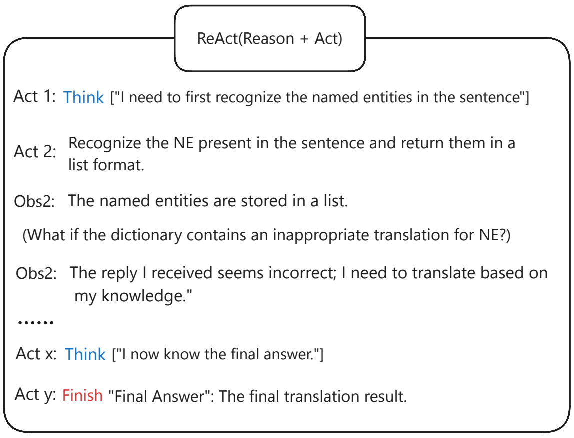
the model's parameters, enhancing its ability to translate in context and maintain named entities as required. Prompts were designed to guide the model step by step using the stored named entity dictionary to replace the entities in the sentence. A few examples and explanations were provided to assist the model in executing the task more effectively. After the adjustments, the M-ETA score significantly improved, while the comet score remained stable. The approach utilized two LLMs from the same series. It raises the question of whether a single model could be used with different prompts designed for various tasks.

In this method, the following considerations were made: for the languages IT and TR, the model underperformed on the comet metric, possibly due to issues within the translation module. Similarly, the model's performance on the M-ETA metric was suboptimal for TR and ZH-TW, likely stemming from the named entity recognition module. It is being considered whether fine-tuning the corresponding modules on specific languages can improve the model's performance for those languages.

### 3.4 Method 4: Qwen with Reason and Act

To further explore the capabilities of LLMs, the combination of reasoning and execution abilities in translation tasks involving named entities is investigated. It examines whether the model can effectively handle unexpected situations, such as the absence of corresponding translations in the named entity dictionary, while enhancing human interpretability and translation reliability. The ReAct method (Yao et al., 2023), based on the LangChain framework, is employed to implement the translation task.

In this method, the high generalization ability of the LLM for translation tasks is validated. The model can quickly understand the task requirements and sequentially generate the translation steps by providing only a few examples in the sample. Especially when multiple potential named entities are present in a sentence, the model first identifies them and produces a list of named entities. It then uses a tool function to search for corresponding translations. If no translation is found or the result is deemed unsatisfactory, the model automatically calls a new tool function to search for better matches. The model can theoretically achieve more accurate named entity translation by constructing a feasible tool function and providing effective information sources.


ReAct(Reason + Act)

Act 1: Think ["I need to first recognize the named entities in the sentence"]

Act 2: Recognize the NE present in the sentence and return them in a list format.

Obs2: The named entities are stored in a list.

(What if the dictionary contains an inappropriate translation for NE?)

Obs2: The reply I received seems incorrect; I need to translate based on my knowledge."

......

Act x: Think ["I now know the final answer."]

Act y: Finish "Final Answer": The final translation result.

Figure 2: Combining ReAct with LLM approach

| System | ar_AE | de_DE | es_ES | fr_FR | it_IT | ja_JP | ko_KR | th_TH | tr_TR | zh_TW | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method 1 | 47.10 | 60.92 | 66.71 | 57.67 | 68.66 | 47.95 | 41.62 | 37.99 | 59.58 | 53.81 | 54.20 |
| Method 2 | 75.23 | 71.54 | 76.09 | 76.19 | 70.09 | 75.78 | 72.52 | 69.51 | 70.00 | 71.88 | 72.88 |
| Method 3 | **91.47** | **88.09** | **91.72** | **90.04** | **92.35** | **91.54** | **91.41** | **89.84** | **86.05** | **86.91** | **89.94** |
| Method 4 | 89.38 | 86.62 | 90.12 | 89.22 | 91.13 | 89.27 | 90.48 | 88.68 | 83.87 | 85.51 | 88.43 |

Table 1: Results of different methods on the task across different languages. Language codes: Arabic (AR), German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Korean (KO), Thai (TH), Turkish (TR), and Chinese (ZH).

| | ar_AE | | de_DE | | es_ES | | fr_FR | | it_IT | | ja_JP | | ko_KR | | th_TH | | tr_TR | | zh_TW | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **System** | M | C | M | C | M | C | M | C | M | C | M | C | M | C | M | C | M | C | M | C | M | C |
| Method 1 | 82.47 | 32.96 | 83.28 | 48.02 | 85.54 | 54.67 | 82.46 | 44.34 | 85.74 | 57.26 | 83.64 | 33.61 | 82.88 | 27.79 | 71.10 | 25.92 | 85.34 | 45.77 | 78.29 | 40.99 | 82.07 | 41.13 |
| Method 2 | 91.06 | 64.09 | 88.24 | 60.16 | 87.57 | 67.27 | 91.41 | 65.31 | 75.51 | 65.40 | 92.47 | 64.19 | 91.84 | 59.92 | 90.65 | 56.37 | 88.64 | 57.84 | 91.54 | 59.17 | 88.89 | 61.97 |
| Method 3 | **94.33** | **88.78** | **94.38** | **82.59** | **95.28** | **88.42** | **93.77** | **86.59** | **94.96** | **89.88** | **95.68** | **87.74** | **94.90** | **88.17** | **93.43** | **86.51** | **94.09** | **79.28** | **94.24** | **80.64** | **94.51** | **85.86** |
| Method 4 | 93.78 | 85.33 | 94.21 | 80.17 | 94.69 | 85.97 | 93.47 | 85.34 | 94.31 | 88.15 | 94.29 | 84.76 | 93.12 | 87.99 | 92.76 | 84.94 | 93.27 | 76.19 | 94.12 | 78.35 | 93.80 | 83.72 |

Table 2: Results across languages with M-ETA (M) and Comet (C) scores. Language codes: Arabic (AR), German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Korean (KO), Thai (TH), Turkish (TR), and Chinese (ZH).

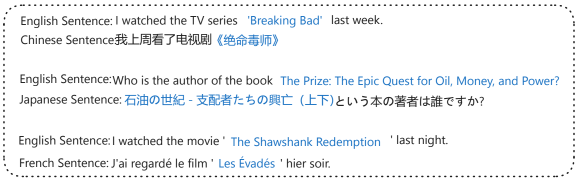
# 4 Results and Analysis

## 4.1 Dataset

This section introduces the dataset used in Task 2, which consists of training data, validation data, and predictions provided by GPT-4o and GPT-4o-mini (Conia et al., 2024). Both the training and validation datasets are provided in JSON format. The training data covers six languages: Arabic (ar), German (de), Spanish (es), French (fr), Italian (it), and Japanese (ja). Each data entry includes a unique ID, source language label (uniformly set to English), target language label, source sentence, target sentence, the source of the named entity, and the Wikidata ID. Approximately 5,000 data entries are provided for each language. In addition to the six languages mentioned, the validation data includes four additional languages: Korean (ko), Thai (th), Turkish (tr), and Traditional Chinese (zh-TW). Furthermore, each validation data entry includes all fields from the training data, along with the named entity type (e.g., "Artwork," "Movie"). Each validation entry may contain multiple translation versions. While the scale of validation data for each language is relatively small, it exhibits high diversity in language types and task complexity.

## 4.2 Evaluation Metrics

This section primarily introduces the evaluation metrics used in Task 2, which assess the translation of named entities (NER) and the overall sentence translation. The final evaluation formula is shown in Equation 2.


English Sentence: I watched the TV series 'Breaking Bad' last week.
Chinese Sentence我上周看了电视剧《绝命毒师》

English Sentence:Who is the author of the book The Prize: The Epic Quest for Oil, Money, and Power?
Japanese Sentence: 石油の世紀 - 支配者たちの興亡（上下)という本の著者は誰ですか?

English Sentence:I watched the movie ' The Shawshank Redemption ' last night.
French Sentence: J'ai regardé le film ' Les Évadés ' hier soir.

Figure 3: Example sentences from dataset

### 4.2.1 COMET

COMET (Cross-lingual Optimized Metric for Evaluation of Translation) is a metric used to evaluate the quality of machine translation systems. It is based on comparing the output of a machine translation system to the output of a human translation system. COMET uses a pre-trained model to generate a score for each translation, which is then used to evaluate the quality of the translation.

$$Sim = 0.2 * N + 0.3 * M + 0.5 * K \tag{1}$$

### 4.2.2 M-ETA

M-ETA (Manual Entity Translation Accuracy) is a metric used to evaluate the accuracy of entity translation in machine translation systems. At a high level, given a set of gold entity translations and predicted entity translations, M-ETA computes the proportion of correctly translated entities in the predicted entity translations.

### 4.2.3 Overall Score

The final evaluation score will be the harmonic mean of the two scores, i.e.:

$$overall = 2 * \frac{M - ETA * COMET}{M - ETA + COMET} \tag{2}$$

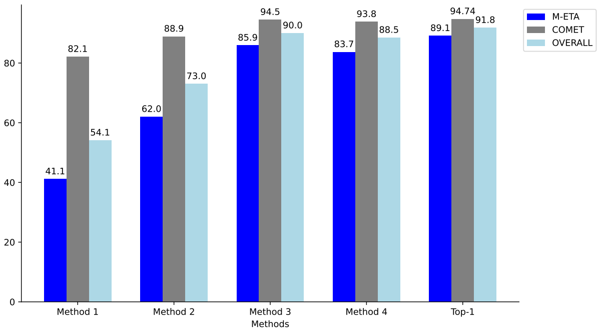


Figure 4: Our four methods compared with the Top-1.

### 4.3 Implementation Details

This section primarily discusses the details of the implementation of the parameters. When calculating the weighted similarity, experiments showed that the most accurate results and the highest M-ETA score were achieved when the statements, site links,and string ontology similarity weights were set to 0.2, 0.3, and 0.5, respectively. As shown in Formula 1, where $N$ represents the number of statements, $M$ represents the number of site links, and $K$ represents the similarity calculated between two named entity strings using the Levenshtein Distance algorithm. In the regularization and cleaning of named entities, the focus is on potential brackets, excess punctuation, and special characters, without altering their character encoding (since languages like Arabic and Chinese may have multiple writing systems, the original version is retained without modification). In the named entity recognition section, the Qwen model does not involve temperature or top-p parameters adjustments. In the MT translation section, multiple experiments showed that setting the temperature parameter to 0.75 resulted in stable M-ETA values while maximizing the comet score.

## 5 Results

In the SemEval 2025 Task 2: EA-MT, the system performed translations for all 10 languages. The model ranked fifth without using Wikidata IDs or information, fifth without RAG, and seventh without fine-tuning. The final overall ranking was 11th. As shown in Table 1, the final scores for each language were presented for the four methods. It can be observed that the method combining two LLM models achieved the best result. Although the React method, implemented using the LangChain framework, yielded slightly lower results, it significantly increased human interpretability and the credibility of the outcomes, facilitating subsequent debugging and improvement. The COMET and M-ETA and overall scores for each method are shown in Table 3. The details of each language are shown in the Table 1 and Table 2.

| | Average across all languages | | |
|---|---|---|---|
| **System** | **M-ETA** | **Comet** | **Overall** |
| Method 1 | 41.133 | 82.074 | 54.081 |
| Method 2 | 61.972 | 88.893 | 73.031 |
| Method 3 | **85.860** | **94.506** | **89.976** |
| Method 4 | 83.719 | 93.802 | 88.474 |

Table 3: Results of four methods on the task.

## 6 Conclusion

This study implements four methods for accurately translating sentences containing named entities in this shared task. A translation method based on traditional MT models was developed, offering speed, ease of deployment, and a certain level of named entity recognition. Multiple LLMs were combined for models based on LLMs, focusing on named entity recognition and translation. A named entity dictionary was utilized to improve the translation accuracy of entities. Additionally, the LangChain framework was used to test the performance of LLMs in generating reasoning trajectories and task-specific actions in the translation of named entities. Our experiments show that, with an additional named entity dictionary, LLMs can be effectively used for cross-lingual content translation involving unknown or complex named entities. For future work, the tool functions in LangChain to assist LLMs in translation, and it is expected that fine-tuning and RAG techniques will be combined to enhance the performance of LLMs in EA-MT further.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

## References

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. Ustc-nelslip at semeval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity recognition. *arXiv preprint arXiv:2203.03216*.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

Xiaojing Du, Hanjie Zhao, Danyan Xing, Yuxiang Jia, and Hongying Zan. 2024. Mrc-based nested medical ner with co-prediction and adaptive pre-training. *arXiv preprint arXiv:2403.15800*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Ichiro Ide, Reiko Hamada, Shuichi Sakai, and Hidehiko Tanaka. 2003. Compilation of dictionaries for semantic attribute analysis of television news captions. *Systems and Computers in Japan*, 34(12):32–44.

Zhenfei Ju, Jian Wang, and Fei Zhu. 2011. Named entity recognition from biomedical text using svm. In *2011 5th international conference on bioinformatics and biomedical engineering*, pages 1–4. IEEE.

Kuanghong Liu, Jin Wang, and Xuejie Zhang. 2023. Entity-related unsupervised pretraining with visual prompts for multimodal aspect-based sentiment analysis. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, pages 481–493.

Ruixuan Lu, Zihang Tang, Guanglong Hu, Dong Liu, and Jiacheng Li. 2023. Netease. ai at semeval-2023 task 2: Enhancing complex named entities recognition in noisy scenarios via text error correction and external knowledge. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 897–904.

Xiang Luo, Jin Wang, and Xuejie Zhang. 2022. Ynu-hpcc at rocling 2022 shared task: A transformer-based model with focal loss and regularization dropout for chinese healthcare named entity recognition. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 335–342.

Chen Lyu, Bo Chen, Yafeng Ren, and Donghong Ji. 2017. Long short-term memory rnn for biomedical named entity recognition. *BMC Bioinformatics*, 18:1–11.

Xinge Ma, Jin Wang, and Xuejie Zhang. 2021. Ynu-hpcc at semeval-2021 task 11: Using a bert model to extract contributions from nlp scholarly articles. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 478–484.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional lstm-crf for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, pages 531–540.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 548–554.

Sujan Kumar Saha, Pabitra Mitra, and Sudeshna Sarkar. 2008. Word clustering and word selection based feature reduction for maxent based hindi ner. In *Proceedings of ACL-08: HLT*, pages 488–495.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint arXiv:cs/0306050*, pages 142–147.

Sergey Vychegzhanin and Evgeny Kotelnikov. 2019. Comparison of named entity recognition tools applied to news articles. In *2019 Ivannikov Ispras Open Conference (ISPRAS)*, pages 72–77. IEEE.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*, pages 1–9.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–33.