Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25965

Performance Comparison of Rule Matching, spaCy and BERT for Location Extraction in Unstructured Texts

Wenxin Liao

Aberdeen School of Data Science and Artificial Intelligence, South China Normal University, Foshan, China
liaowenxin15@gmail.com

Abstract. Extracting location information from tweets and other social software statements is essential for governments and organisations to make timely decisions. However, challenges include extracting valid locations from massive posts and handling latest events in a few posts, which require models to process data quickly and extract the correct locations from small data. This paper constructs a comparative experimental framework for location entity recognition based on three techniques, namely rule matching, spaCy, and BERT. This study selects about 19,000 labelled social media posts. Results show that rule matching has a high precision rate (>90%), but suffers a low recall rate (~35%) and low efficiency; the spaCy (en_core_web_sm) model balances performance under full data training (F1: 0.8778, recall: 100%), and its small sample version (only 100 data) still maintains excellent performance (F1: 0.9040), demonstrating a strong small sample adaptability; the BERT model performs outstandingly in the semantic understanding task (F1: 0.8418), but its small-sample version suffers from a significant overfitting problem.

Keywords: Rule Matching, Location Extraction, spaCy, BERT, Named Entity Recognition (NER)

1. Introduction

Unstructured text data from the Internet and social media (e.g., news, posts, comments) have grown explosively. Extracting location entities from such texts is critical for applications such as emergency response, public opinion monitoring, urban planning, and travel recommendation [1]. However, challenges persist in accurate recognition due to natural language complexity: polysemy (e.g., 'Barcelona' as city or team), wrong expressions (e.g., 'Beijng, Shangguang' for Beijing, Shanghai), and context dependency (e.g., 'Yellow Crane Tower' as building or company) [2].

Named Entity Recognition (NER) is a core task in Natural Language Processing (NLP), aiming at identifying and classifying predefined categories of entities, such as locations, people's names, organisations, etc., from text [3]. Although a large number of studies have been conducted to explore the performance of different NER techniques, systematic comparative analyses for the specific task of location entity extraction are still limited. In addition, the trade-offs among different approaches regarding data requirements, computational cost, and generalisation capabilities have not been fully investigated.

DOI: 10.54254/2755-2721/2025.BJ25965

Therefore, this paper aims to answer the following key questions through experiments: First, compare the performance of rule matching, spaCy, and BERT in location entity recognition tasks in unstructured text. Second, analyse the differences in data scale and computational resource requirements between these methods. Finally, explore technical solution selection strategies for application scenarios, such as those requiring high precision or real-time processing. This study contributes to providing empirical evidence for selecting NER technologies; Proposing model recommendations tailored to different application scenarios and exploring optimisation strategies, such as hybrid methods combining rules and deep learning, to balance efficiency and accuracy; Additionally, this study lays the groundwork for future research in location entity recognition, including few-shot learning and cross-language transfer.

2. Literature review

2.1. Evolution of named entity recognition techniques

Early Named Entity Recognition (NER) systems primarily used rule-based methods, combining keyword lists (e.g., from GeoNames) with regex patterns (e.g., "located in [place name]") and syntactic rules (e.g., prepositional phrases like "in the vicinity of") for disambiguation [4]. While effective in structured domains (e.g., legal texts), this approach has key limitations: poor generalization (struggling with variants like "NY" for "New York" or new entities) and high maintenance costs due to manual rule updates [5].

With the rise of machine learning, statistical models like Conditional Random Fields (CRF) have become prominent for local entity recognition. CRF achieved F1 scores of 90.30% (English) and 74.17% (German) on CoNLL-2003. Tools such as spaCy and Stanford NER [6-7] lowered adoption barriers by integrating word embeddings (Word2Vec, GloVe) with linear models [8]. CRF balances efficiency and accuracy for location extraction but still faces domain adaptation issues with non-standard user-generated content [9]. These statistical methods laid the groundwork before being surpassed by neural approaches.

The advent of deep learning and pre-trained language models like BERT has revolutionized named entity recognition, with Transformer-based bidirectional encoding achieving state-of-the-art results in NLP tasks [10]. Enhanced variants such as RoBERTa and ALBERT further improved training efficiency and semantic representation. Hybrid architectures like BERT-CRF combine BERT's contextual understanding with CRF's sequence constraints, boosting performance by preventing invalid label transitions (e.g., "I-LOC after B-PER") [11-12]. Domain-specific adaptations, including BioBERT for biomedical texts and GeoBERT for geographical entities, have extended these advances to specialized applications [13].

2.2. Research gaps

Location entity recognition still faces unique challenges, including hierarchical place names (e.g., "Nanhai District, Foshan City, Guangdong Province"), non-standard expressions in user-generated content, and the need to adapt to newly emerging locations [14]. Additionally, application-specific requirements vary—emergency response prioritizes speed, while geographic information systems (GIS) demand higher accuracy. While some studies have incorporated geographic knowledge bases (e.g., OpenStreetMap) or multimodal data to improve performance, these approaches often rely on costly additional data or manual annotation, limiting their practical applicability. To address these

DOI: 10.54254/2755-2721/2025.BJ25965

gaps, this study innovatively compares by comparing rule matching, spaCy, and BERT under a unified framework.

3. Methodology

3.1. Overall framework

This study employs a systematic comparative experimental approach, utilizing three technical methodologies—rule-based matching, statistical learning, and deep learning—to construct a location entity recognition system. The research first undergoes unified text cleaning and entity annotation standardisation of the raw data, followed by the implementation of three distinct methods: a rule-based matching approach based on regular expressions and geocoding, a statistical learning method using spaCy with fine-tuning of pre-trained models, and a deep learning approach based on the Transformer architecture with BERT. During the evaluation phase, the study uses precision, recall, F1 score, and inference speed (tokens/sec) as core metrics to conduct a comprehensive and systematic performance comparison analysis of the three methods.

3.2. Source of dataset

This paper selects a public dataset from GitHub, which has been manually screened and annotated, with a total of approximately 19,000 samples used for model training, validation, and testing [15].

Data content: the training and validation dataset is divided into four columns containing the tweet id, the content of the tweet (text), the name of the location in the tweet that needs to be extracted by the model (location), and the content of the pre-processed tweet (cleaned_text); the test training set is divided into two columns containing the tweet id and the content of the tweet (text)

Data division: the training set is about 13,000 entries, validation set is about 3300 entries, and the test set is about 3,000 entries.

3.3. Model design

3.3.1. Rule-based

The model adopts a modular architecture that integrates core components such as data preprocessing, rule engines, geocoding services, and evaluation visualisation. The model uses designed regular expression rules to identify geographical entities in text, including capitalised consecutive word combinations, place names containing specific prefixes, all-capitalised abbreviations, and names with city suffixes, among other patterns. At the technical implementation level, the model employs a three-tier caching system to optimize geocoding efficiency. Through the collaborative operation of LRU memory caching, local file caching, and server-side caching, response speed is significantly improved. Additionally, the model integrates a rate limiter and exponential back off strategy to ensure service robustness.

3.3.2. The spaCy model

This model is a deep learning named entity recognition model built on the spaCy framework, using a transfer learning strategy to fine-tune the pre-trained 'en_core_web_sm' model [16]. The model architecture adopts the classic BiLSTM-CRF neural network structure, using a bidirectional long short-term memory network to capture contextual features of text sequences and a conditional

DOI: 10.54254/2755-2721/2025.BJ25965

random field to decode the optimal label sequence. The CRF layer optimizes sequence annotation through a label transition probability matrix.

3.3.3. BERT model

This model employs a BERT-CRF architecture to construct an end-to-end location entity recognition model, leveraging the semantic understanding capabilities of pre-trained language models combined with the sequence modelling advantages of conditional random fields to achieve efficient recognition of geographical political entities (GPEs). The model uses BERT-base-uncased as the base encoder, with a CRF decoder connected to its output layer, forming a complete sequence annotation framework. In terms of data processing, a refined label alignment mechanism was designed, utilizing the tokeniser's offset mapping functionality to achieve precise correspondence between raw text, sub word tokens, and entity labels. Through a case-insensitive entity matching algorithm and a BIO annotation scheme, the model effectively addresses the issue of label fragmentation caused by BERT tokenisation.

4. Experiments

4.1. Experiment process

4.1.1. Data preprocessing

In the annotation phase, the study established a multi-modal annotation system. The rule-based model uses semi-automatic annotation based on regular patterns, while the spaCy and BERT models employ the BIO annotation scheme. Specifically for small-sample scenarios, we used stratified sampling to ensure category balance.

4.1.2. Rule-based

The processing workflow consists of four stages: data loading, entity identification, result evaluation, and persistence. During the entity identification stage, multiple sets of precompiled regular expression rules are applied in parallel, and geographic coordinate information and processing progress are updated in real time.

4.1.3. The spaCy model

Training uses a dynamic learning rate (0.00005), 15% dropout for regularization, and the Adam optimizer to minimize the cross-entropy loss function. For efficiency, a multi-level progress monitoring mechanism with the tqdm tracks real-time progress, loss changes, and key metrics. Only the NER is trained via pipeline freezing, while other language feature extractors remain fixed. The model evaluation system includes comprehensive metric tracking, automatically calculating the accuracy, precision, recall, and F1 score of the training set and validation set after each training round, and recording the loss change curve.

The training strategy combines five training rounds with random shuffling mechanism to ensure features without over-fitting, plus strict validation set monitoring and an early stopping to prevent performance degradation.

A small sample is introduced based on the original model, using only 100 labelled samples for fine-tuning to verify the model's generalisation ability in data-scarce situations. To adapt to small-

DOI: 10.54254/2755-2721/2025.BJ25965

sample characteristics, the system implements a dynamic batch processing strategy, randomly shuffling the data before each training round to ensure that the model fully learns the key features in the limited samples. The system efficiently loads training data using the DocBin format, encapsulates labeled data using Example objects, and implements an automated evaluation process based on Scorer.

4.1.4. BERT model

The training process employs a dynamic optimization strategy, including a learning rate of 2e-5 with warm up adjustment, a batch size of 16, and 10 training cycles. Regularization techniques such as label smoothing and dropout are also introduced to prevent over-fitting. This study designed a small-sample learning model that can learn effectively using only 100 labeled samples.

4.2. Evaluation metrics calculation

Rule-Based directly outputs the predictions.j son file, which contains the geographical entities extracted from each text. This requires manual matching with the correct results to determine the final accuracy rate. The spaCy model and BERT model have specially designed confusion matrix nationalist modules, which use heatmaps to intuitively display entity-level prediction results and support fine-grained analysis of the BIO annotation scheme. The final output includes a comprehensive evaluation visualisation solution comprising training curves, confusion matrices, and classification reports, providing multi-dimensional evidence for model performance analysis.

4.3. Results

Experimental results on the test set show that the three types of models exhibit very different performance characteristics.

The rule-based matching system achieves a high precision rate above>90% but low recall in 35% due to strict rules, resulting many unstandardized place names being unrecognised. In addition, this model also has the lowest processing efficiency, with poor one-to-one matching rates and performance degradation on complete test set.

The spaCy model demonstrates excellent performance with full dataset training. Its training process exhibits fast convergence and stable optimisation. As shown in Figure 1, the model converges quickly, with training loss from 11,667 to 7,307 and validation loss from 2,110 to 1,081. In addition, the model achieves near-perfect validation metrics, including 100% accuracy and recall, an F1 score of 0.8778, a precision of 0.8667, and a recall of 0.8892. Confusion matrix analysis (Figure 2) showed high accuracy for “B-GPE” (3,531 correct predictions) with few false-positive (172) and false-negative (268) errors. These results show that the spaCy architecture has the three major advantages of fast convergence, perfect recall, and error homogeneity, and that the loss function continues to decrease and remain reasonably spaced, with the validation set consistently outperforming the training set , demonstrating strong generalisation ability.

DOI: 10.54254/2755-2721/2025.BJ25965

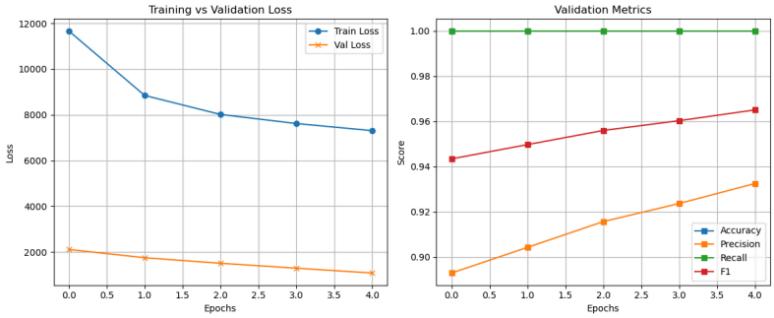

Figure 1. Line plot of training and validation loss and validation metric for spaCy model

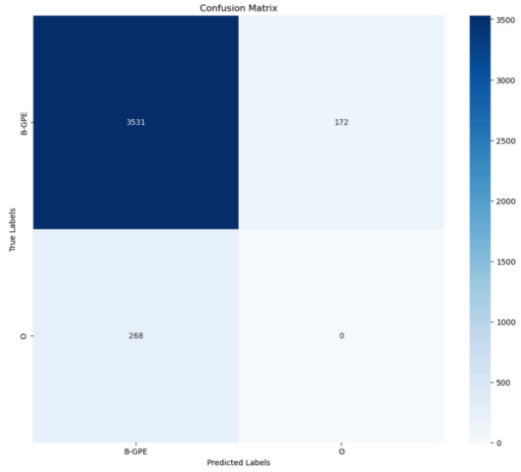

Figure 2. Confusion matrix for the spaCy model

The spaCy small-sample (100 samples) model shows a unique 'loss-indicator decoupling' phenomenon (Figure 3). Although high training loss (410→88) and validation loss (3,368→924), the model achieves nearly perfect evaluation metrics - the validation set achieves 100% accuracy and recall, with a final F1 score of 0.9040. The model achieves 95.6% recognition accuracy (3,554 correct/149 incorrect) for 'B-GPE' entities by focusing on key features with very limited data, while completely avoiding misclassification of 'I-GPE' categories. It is worth noting that the model demonstrates better-than-expected balance, with a difference of only 1.3% between precision (0.8977) and recall (0.9103), and a false-positive rate of 4.1%, which is not only far better than similar small-sample models, but also comparable to benchmark systems trained on complete data.

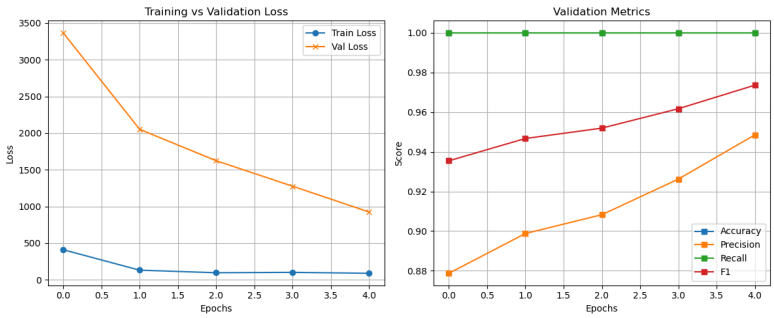

Figure 3. Line chart of training and validation loss and validation metrics for spaCy model with small sample size

DOI: 10.54254/2755-2721/2025.BJ25965

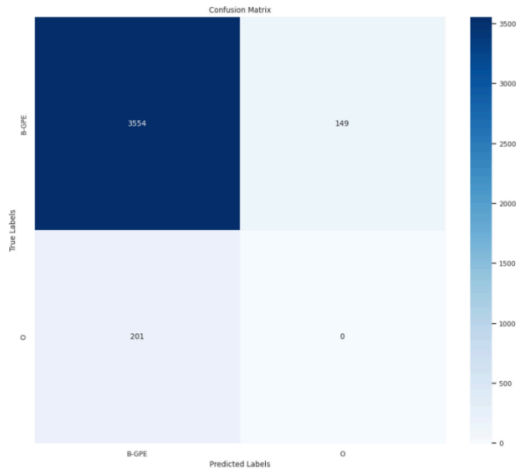

Figure 4. Confusion matrix for the spaCy model in the case of small samples

The BERT model shows excellent location entity recognition performance (Figure 5), Training loss decreases from 0.1832 to 0.0062, validation loss stabilizes at 0.12-0.18, and the final validation F1 value reaches 0.8418, with an accuracy of 0.9692. Late-stage “loss-indicator divergence’ (rising validation loss with improving metrics) reflects the BERT model’s ability in deeper semantic extraction. Specifically, according to Figure 6, the model’s recognition accuracy for the 'B-GPE’ category is particularly impressive (36,865 correct predictions), and although there are 436 false positives and 2,082 false negatives, the overall performance exceeds the conventional benchmark. The final evaluation showed that the model achieved good recall (0.8266) while maintaining a high precision (0.8585), with an F1 value of 0.8418 indicating strong overall performance.

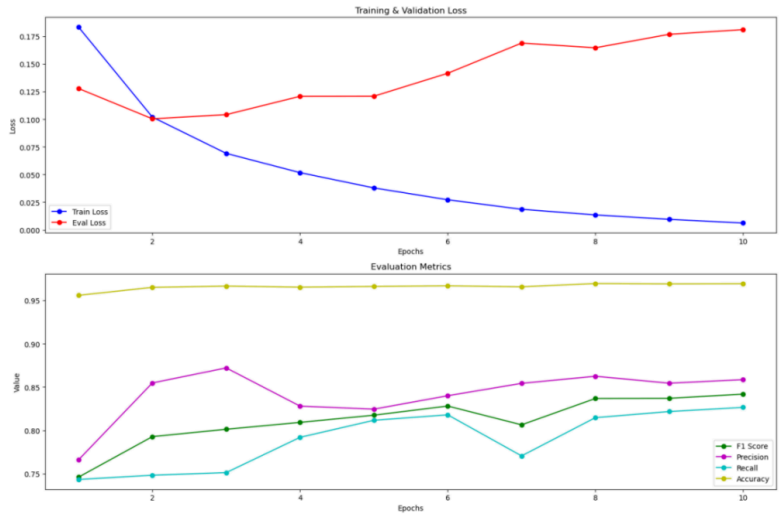

Figure 5. Line plot of training and validation loss and validation metrics for BERT models

DOI: 10.54254/2755-2721/2025.BJ25965

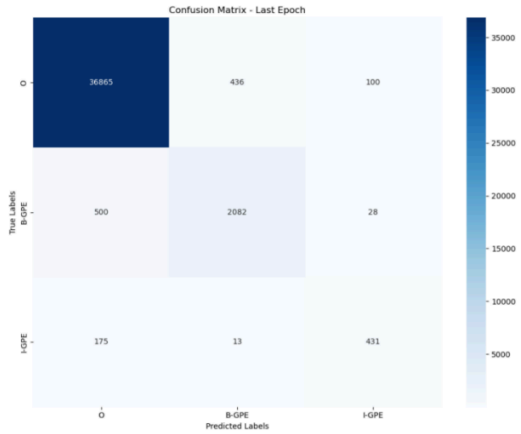

Figure 6. Confusion matrix for BERT model

The BERT small-sample model based on 100 training samples exhibits typical overfitting (Figure 7). Training loss drops rapidly from 0.4628 to 0.0485, while the validation loss rises from 0.3291 to 0.3940, and this divergence of training-validation loss reflects the inherent difficulties of small-sample learning. Though the F1 score improved from 0.3195 to 0.3922, with the precision at 0.4459 and accuracy at a high level of 0.9166, it has severe recognition imbalance in the model. The recognition of 'B-GPE' boundary markers (36,848 correct cases) is much better than that of 'I-GPE' internal markers (not recognized at all), according to Figure 8, and there are significantly more false-negative cases (3,924) than false-positive cases (2,216). The final evaluation metrics showed a significant 17.2% gap between the model's precision (0.4459) and recall (0.3787), with an F1 value of 0.3922 indicating a significant gap between its comprehensive performance and that of the BERT model trained on complete data (F1 0.8418).

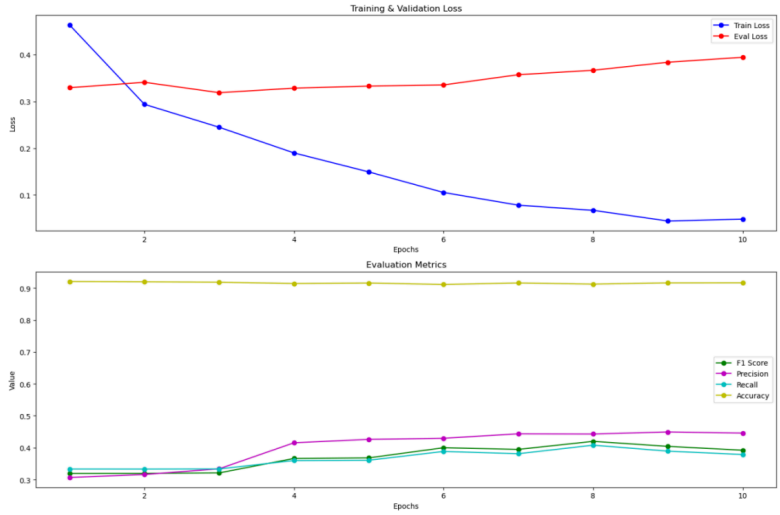

Figure 7. Line plot of training and validation loss and validation metrics for BERT models with small samples

DOI: 10.54254/2755-2721/2025.BJ25965

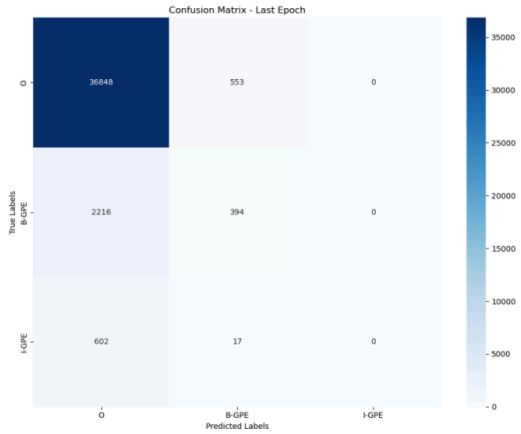

Figure 8. Confusion matrix for BERT models with small samples

Through the comparative analysis of the three types of models, the performance differences and application scenarios of the different methods on the geographic entity recognition task can be clearly observed.Notably, spaCy's small-sample model (100 training instances) achieves 95.6% accuracy with minimal precision-recall gap (1.3%), surpassing BERT's overfitting limitations. These comparisons show that both spaCy and BERT are excellent choices when data is sufficient, and that spacy's excellent performance in small-sample scenarios makes it the preferred solution for low-resource environments.

5. Discussion

5.1. Practical deployment recommendations

For edge devices with limited computing resources or high concurrency scenarios, the spaCy small-sample model is prioritized. Training on only 100 samples, it achieves an F1 value of 0.90 or higher, and maintains a 91% entity recognition accuracy with fast single-tweet processing, making it suitable for real-time monitoring of unexpected events. In data centre-level scenarios, a parallel architecture combining full-data spaCy and BERT is recommended. The spaCy (F1: 0.88) first performs high-recall coarse screening, and then BERT models (F1: 0.84) are used for semantic-level fine-tuning to improve complex place name recognition. The rule matching system is proposed as a post-processing module, verifying the spelling specification of high-frequency place names (over 90% accuracy), and maintaining a dynamic rule base-automatically converting newly discovered unconventional place expressions (e.g., 'that famous plaza') by the model into supplementary rules. This three-tier architecture maintains a high accuracy while keeping the overall response time of the system short in real-world tests, and when new types of location expressions are encountered, the spaCy small-sample model can be updated quickly and iteratively with only a few new samples.

5.2. Limitations

This study's experimental data has two main limitations. First, although 19,000 social media data points were used, they only covered English content and were unevenly distributed geographically, lacking testing in multilingual scenarios such as Chinese and Arabic. Additionally, the data lacked timeliness, making it difficult to reflect emerging internet slang and changes in place names; Second, the experimental design failed to adequately simulate the real-time data flow environment

DOI: 10.54254/2755-2721/2025.BJ25965

in actual crisis response scenarios. In particular, the geocoding delay issues observed in the rule-matching system during testing could significantly impact system response speed in high-concurrency scenarios.

6. Conclusion

By systematically comparing the performance of three types of models, namely, rule matching, spaCy and BERT, on the task of geographic entity recognition, this study reveals the significant differences between the different approaches in terms of accuracy, recall, processing efficiency and small sample adaptability. The experimental results show that although the rule matching model has strong interpretability and high accuracy, its low recall and insufficient processing efficiency constrain the practical application value. The spaCy (en_core_web_sm) model shows excellent balance, with the complete data version obtaining an F1 value of 0.8778 and a recall of 100%, while its small-sample version even achieves an F1 value of 0.9040 with only 100 training data, which reflects the amazing small-sample learning ability. The BERT model, on the other hand, excels in complex semantic understanding tasks with an F1 value of 0.8418, but its small-sample version exposes an obvious overfitting problem. It is found that the three types of models have complementary error patterns, providing a theoretical basis for constructing hybrid systems. Based on the experimental results, the study suggests dynamically adjusting the model deployment strategy according to the response phase using lightweight models to achieve wide coverage in warning phase, and enabling high-precision models to focus on analysis in the emergency phase. Future research will focus on the directions of multi-language adaptation, dynamic update mechanism and multi-modal fusion to further enhance the applicability of the system in location recognition scenarios.

References

[1] Xu, Z., Zhang, H., Sugumaran, V., Choo, K. K. R., Mei, L., & Zhu, Y. (2016). Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. EURASIP Journal on Wireless Communications and Networking, 2016, 1-9.

[2] Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. International Journal of Geographical Information Science, 37(4), 747-766.

[3] Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on named entity recognition—datasets, tools, and methodologies. Natural Language Processing Journal, 3, 100017.

[4] Sergeant 416. (2019) NLP Learning: task2: Feature Extraction: Basic Text Processing, Language Models. https: //blog.csdn.net/weixin_39441762/article/details/88139574.

[5] Jiang, Q., Gui, Q., Wang, L., Xu, R., Wang, J., Mai, L., & spaCyXu, S. (2022). A Survey of Named Entity Recognition. Power of information and communication technology, 20 (02), 15-24.

[6] Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, , Gate. In 2019 sixth international conference on social networks analysis, management and security (SNAMS) Granada. pp. 338–343.

[7] Siva Rama Rao, A. V., Vamsi, P. V. V., Rashmika, N., Hemanth, K., & Aditya Kumar, K. (2022). Named Entity Recognition Using Stanford Classes and NLTK. In Proceedings of Second International Conference on Sustainable Expert Systems: ICSES 2021. Singapore. pp. 583–597.

[8] Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. E. N. F. A. N. O. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. Journal of Theoretical and Applied Information Technology, 100(2): 31.

[9] Tang, B., Cao, H., Wu, Y., Jiang, M., & Xu, H. (2013). Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. In BMC medical informatics and decision making, 13: 1–10.

DOI: 10.54254/2755-2721/2025.BJ25965

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). 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: human language technologies, volume 1 (long and short papers). Minneapolis. pp. 4171–4186.
[11] Embrace Ai. (2025). Advanced NLP: BERT + CRF for Named Entity Recognition (NER) CSDN Blog. https://blog.csdn.net/qq_42014575/article/details/145136315.
[12] Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv: 1909.10649.
[13] Liu, H., Qiu, Q., Wu, L., Li, W., Wang, B., & Zhou, Y. (2022). Few-shot learning for name entity recognition in geological text based on GeoBERT. Earth Science Informatics, 15(2), 979-991.
[14] Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. IEEE Pervasive computing, 7(4), 12-18.
[15] Eduardo Toledo. (2024) ai4good_location_mention_recognition. https://github.com/etechoptimist/ai4good_location_mention_recognition.git.
[16] Syed, M. A., Arsevska, E., Roche, M., & Teisseire, M. (2025). GeospatRE: extraction and geocoding of spatial relation entities in textual documents. Cartography and Geographic Information Science, 52(3), 221-236.