Natural Language Processing (CS-354)

Literature Review

Session: 2020 – 2024

Submitted by:

Muhammad Zeshan Ayyub	2020-CS-113
Muhammad Ali Murtaza	2020-CS-114
Ali Tariq	2020-CS-142
Syed Azeem Ali Hashmi	2020-CS-156

Supervised by:

Prof. Dr. Muhammad Usman Ghani Khan

Department of Computer Science

University of Engineering and Technology

Lahore Pakistan

Contents

List of Tables **ii**

1 Literature Review 1

List of Tables

1 Literature Review

Like other sub-tasks of NLP research, the early research of NER mainly focused on rules and dictionaries. The design of rules or dictionaries was generally based on syntax, grammar, lexicon patterns, and knowledge in specific fields. Puccetti et al. (2022) [11] compare different methods to get the best results. The targeted problem is to extract the technological terms from the patent text to forecast future technologies. The database chosen for retrieving the patent documents is the Erre Quadro. The database contains over 90 million patents and includes high quality bibliographical and legal status patent information from leading industrialized and developing countries. The techniques applied includes Gazetteer, Rule Based and Deep Learning (DL) based (e.g. BERT). The technique further divides this database into 4 different IPC sections depending on the domain of the patent. BERT is the technique that achieve maximum relative recall 0.773 in 3 groups out of 4 that proves BERT the better technique among others.

Luthfi et al. (2022) [4] establish a technique for automated hadith narrator identification using BERT based named entity recognition. This study uses texts from the Bukhari Hadith Book, and a total of one hundred and two Hadith texts in indonasian language are formatted entirely in IOB format. Hadith text first converted to tokens with pretrained model (cahya/bert-base-indonesian-1.5G) then append special tokens of CLS and SEP for model performance. These special tokens then map to their id's and truncate the sentence to max length with constant number. Then create attention mask and pass the mask to BERT for NER. This whole process achieve 98.27 F1 score in correctly recognizing hadith narrator in indosian language.

Lee et al. (2022) [8] introduces NLP technique that can classify named entities from social media text into 6 classes (i.e. Person, Location, Group, Corporation, Product, Creative Work). They used multiple datasets including MSRA (2006), Weibo (2015), People Daily, Boson and CLUENER (2020) to train the system. System used BERT embedding for character representation and then train the BiLSTM-CRF for recognizing complex named entities. Model achieve 0.8468 precision, 0.853 recall and 0.8096 f1 score.

Patil et. al. (2022) [10] present first gold standard named entity recognition dataset in Marathi Language. Dataset consist of 25,000 sentences in the Marathi language. The master dataset used for the compilation of the study is L3Cube-MahaCorpus (2022) which is a monolingual Marathi dataset majorly from news domain. The dataset is tagged in 7 named entity classes (i.e. Person, Location,

Organization, Measure, Time, Date, and Designation). Patil et. al. benchmark different models like CNN, LSTM and Transformer based models like mBERT, XML-RoBERTa, IndicBERT ad MahaBERT. MahaBERT provides 85.30 f1 score, 84.27 precision, 86.36 recall and 97.18 accuracy whcih is the highest evaluation matrix among other models.

Yang et. al. (2022) [14] proposes a new model, the BERT-Star-Transformer-CNN-BiLSTM-CRF, to solve the prblem of computational efficiency of the traditional transformer. The MSRA, Weibo both chinese dataset used in the study. The methodology of this study is devided in three sections Word Vector Embedding, Feature Extraction and Feature Fusion. The study core part word vector embedding using BERT approach to solve the problem of multiple meanings of a word. This contributes to less data to process and result in 40% computational efficiency with the rest of techniques

Berragan et. al. (2022) [3]proposed a technique to extract place names from the given text. Embedding place names in online natural language text represents a useful source of geographic information. They used five custom-built named entity recognition models to perform this task. The data used in this research was from Wikipedia which was extracted from DBPedia. The annotation format of ConLL-03 is used to annotate the data. Other techniques such as DistilBERT BiLSTM CRF Stanza are also specified. Their best-performing model achieves an F1 score of 0.939 compared with 0.730 for the best-performing pre-built model.

Huang et. al. (2022) [7] recognized the problem of extracting accurate information from massive available geological data. They proposed geological news named entity recognition (GNNER) method based on the bidirectional encoder representations from transformers (BERT) pre-trained language model. This solves the problem of traditional word vectors that are difficult to represent. First, they used the method that uses the BERT pre-training model to embed words in the geological news text, and then dynamically obtained word vector is used as the model's input. Second, the word vector is sent to a bidirectional long short-term memory model for further training to obtain contextual features. Finally, the corresponding six entity types are extracted using conditional random field sequence decoding. The data was collected from geological news texts from the China Geological Survey and annotated data using automatic annotation and manual calibration with the open-source annotation tool YEDDA to build a corpus of geological news texts. They achieved an average F1 score of 0.839 identified by the model.

Gorla SK et. al. (2022) [5]felt the need to perform NER on the Telugu language. They performed the NER task in Telugu Language using Word2Vec, Glove, FastText, Contextual String embedding, and bidirectional encoder representations from transformers (BERT) embeddings. They generated the embeddings using Telugu Wikipedia articles from scratch. They used the benchmarked data-set provided by the Forum of Information Retrieval and Evaluation (FIRE-2018). The data consists of 767,603 tokens out of which 200,059 are NEs. Experimental results show that the BERT model outperformed other embedding models as it is trained bidirectionally, taking both the previous and next tokens while predicting and capturing the context of a word. It also captures the syntactic and semantic nature of the language.

Lv X et. al. (2022) [9] proposed a methodology to automatically extract information from Chinese geological reports, namely, geological named entity recognition. They presented Bidirectional encoder representations from transformers (BERT)-(Bidirectional gated recurrent unit network) BiGRU- (Conditional random field) CRF, which is a deep learning-based geological named entity recognition model that is designed specifically with these linguistic irregularities in mind. Based on the pre-trained language model, an integrated deep learning model incorporating BERT, BiGRU and CRF is constructed to obtain character vectors rich in semantic information through the BERT pre-trained language model to alleviate the lack of specificity of static word vectors (e.g., word2vec) and to improve the extraction capability of complex geological entities. The following datasets were used during the study MSRA, Boson, PeopleNER, and GeoNER2021. They achieved an accuracy of 0.724, precision of 0.799, recall of 0.669 and F1-score of 0.728.

Agarwal et. al. (2022) [1] proposes a transfer-learning-based approach for nested named-entity recognition (NER), which outperforms existing machine learning models like conditional random field (CRF) and Bi-LSTM-CRF. It claims that this approach has better generalization capabilities and is simpler compared to existing approaches. It can be applied to other natural language processing tasks as well. Different modeling techniques are employed to solve the problem of nested named-entity recognition, including layering, cascading, and joint labeling. The annotation is performed at multiple levels to capture nested information in the named-entity recognition dataset. Experiment is conducted on these three datasets which contains both nested and flat entities: GermEval 2014, GENIA, and JNLPBA. These datasets contain annotated examples of named entities, including DNA, protein, cell, and more. The details about the number of entity

types, abstracts, sentences, and tokens in each dataset are given respectively: GENIA has 5, 2000, 18,546, 5,60,881, GermEval 2014 has 12, N/A, 31,302, 5,91,005, and JNLPBA has 5, 2404, 24,806, 5,95,994. The experiment is performed using different variants of pre-trained BERT models, Google AI pre-trained BERT model, SciBERT, and BioBERT, for nested NER. They compared the performance of their proposed model with other machine-learning models and existing research work. The techniques used to measure the performance are the F1-score, precision, and recall. These metrics were evaluated using a third-party tool provided during a CoNLL 2000 shared task for evaluating the F1-score. The results show that the pretrained BERT models, with different variants based on domain, size, and cased/uncased versions, perform well in capturing the nested information in the named entities.

Tao et. al. (2022) [12] introduce TPCNER, a large self-annotated corpus of geographic domains with seven categories that are Water System (WAT), Residential land and Facilities (RLF), Landforms (LAN), Organizations (ORG), Transportation (TRA), Pipelines (PIP) and Boundaries, Regions and other areas (BRO) and 64,063 labeled samples. This corpus has more entity categories and larger sample sizes than preceding corpora, demonstrating its efficiency through assessment experimental findings. A novel Chinese NER (CNER) model for the geographic domain is proposed, utilizing the improved ALBERT pretraining model and BiLSTM–CRF. This model enhances overall performance by learning word-level feature representation through the ALBERT layer, extracting text contextual semantic features through the BiLSTM layer, and obtaining the global optimal token sequence through the CRF layer. The performance of ALBERT–BiLSTM–CRF is evaluated by using a range of standard models on various datasets, demonstrating its effective performance on domain-specific and generic datasets. A large-scale annotated corpus, TPCNER, is established with approximately 2 million words from the Baidu Encyclopedia and the Chinese Encyclopedia of Chinese Geography. A new TPCNER annotation tool, ChineseNERAnno, is developed by researchers of this paper to ensure consistency and accuracy in the annotation process. The resulting TPCNER corpus consists of 7 categories, 650,725 entities, and 64,063 samples. The proposed hybrid neural network model for Chinese placename recognition is presented, which includes the ALBERT layer, BiLSTM layer, CRF layer, and output layer. Various experiments and comparisons with other advanced models has been conducted to demonstrate the superior performance of proposed model.

Brandsen et al. (2022) [2] addresses the underutilization of archaeological reports,

often categorized as grey literature, emphasizing the need for advanced search tools to extract valuable information from these documents. By implementing a text retrieval engine and a structured query interface, the study aims to facilitate efficient retrieval of archaeological entities from a large collection of reports in the Netherlands. Named Entity Recognition (NER) is employed to automatically detect and label archaeological entities, enabling users to perform entity-based queries combined with full-text search. The methodology involves indexing the documents, running the inference NER model on each page to detect entities, and storing the entities and full text in a JSON file for each document. The search engine is equipped with a faceted search interface, allowing users to filter results based on document type, subject, geographical location, and date range. The evaluation of BERT models for NER in the Dutch archaeological domain demonstrates the superiority of the domain-specific ArcheoBERT model in achieving the highest F1 score. The study's comprehensive methodology, evaluation, and future work lay a strong foundation for advancing the field of archaeological information retrieval.

Tikayat et. al. (2023) [13] addresses the challenges associated with natural language requirements in aerospace engineering. It emphasizes the need for a model-centric approach to capture the interrelationships of modern complex systems, highlighting the limitations of manual examination processes and the difficulties in verifying requirement completeness. The corpus was collected by gathering aerospace-domain texts, including scientific aerospace texts and requirements from Federal Aviation Requirements (FARs). It contained a total of 1432 sentences. The corpus is annotated using the BIO tagging scheme, and five classes of named entities is identified based on their frequency of occurrence in aerospace texts. Next, the BERT BASE language model (LM) was selected for fine-tuning. BERT BASE was pre-trained on English Wikipedia and BookCorpus. The performance of aeroBERT-NER and BERT BASE -NER was compared in terms of precision, recall, and F1 score. AeroBERT-NER achieved a precision of 0.93, recall of 0.92, and an F1 score of 0.92 whereas BERT BASE-NER was unable to identify any named entities (NEs) apart from two subwords, resulting in lower performance compared to aeroBERT-NER.

He et. al. (2023) [6] focus on a novel approach for Chinese Named Entity Recognition NER using the PWII-BERT model. Experiments are conducted on four benchmark datasets that are Weibo, OntoNotes, Resume and MSRA. The number of sentences in each dataset are 1670, 20,000, 4280, 50,800 respectively. The methodology involves the use of prompt representations, Word-level Information

Injection Adapter (WIIA), and Conditional Random Field (CRF) to enhance the performance of the model. The prompt representations guide the injection of word-level information, while the WIIA module fuses the lexicon features with the original BERT parameters. Fine-tuning is employed to adjust the BERT parameters for better fusion of lexicon features. The CRF layer models the relation between output labels. PWII-BERT outperformed other models on Resume dataset in terms of Precision, Recall, and F1 score, with a Precision score of 95.52, Recall score of 96.87 and an F1 score of 96.19. In the ablation study, individual modules of PWII-BERT were evaluated by removing each component for evaluation. The results showed that the performance dropped by 0.27%, 0.41%, 0.25%, and 0.06% when the CRF decoder, prompt-guided Transformer structure, WIIA component, and lexicon injection were removed, respectively. This indicates that the CRF decoder captures the relationship inside the output tags and improves performance, the category prompt is effective in integrating the lexicon feature, and the lexicon injection contributes to better results.

References

[1] A. Agrawal, S. Tripathi, M. Vardhan, V. Sihag, G. Choudhary, and N. Dragoni. Bert-based transfer-learning approach for nested named-entity recognition using joint labeling. *Applied Sciences*, 12(3):976, 2022.

[2] A. Brandsen, S. Verberne, K. Lambers, and M. Wansleeben. Can bert dig it? named entity recognition for information retrieval in the archaeology domain. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3):1–18, 2022.

[3] A. C. Cillian Berragan, Alex Singleton and J. Morley. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, 37(4):747–766, 2023. doi: 10.1080/13658816.2022.2133125.

[4] T. L. Emha, Z. I. M. Yusoh, and B. M. Aboobaider. Bert based named entity recognition for automated hadith narrator identification. *International Journal of Advanced Computer Science and Applications*, 13(1), 2022.

[5] Gorla. Telugu named entity recognition using bert. *International Journal of Data Science and Analytics*, 2022.

[6] Q. He, G. Chen, W. Song, and P. Zhang. Prompt-based word-level information injection bert for chinese named entity recognition. *Applied Sciences*, 13 (5):3331, 2023.

[7] C. Huang, Y. Wang, Y. Yu, Y. Hao, Y. Liu, and X. Zhao. Chinese named entity recognition of geological news based on bert model. *Applied Sciences*, 12(15), 2022. ISSN 2076-3417. doi: 10.3390/app12157708. URL https://www.mdpi.com/2076-3417/12/15/7708.

[8] L.-H. Lee, C.-H. Lu, and T.-M. Lin. Ncuee-nlp at semeval-2022 task 11: Chinese named entity recognition using the bert-bilstm-crf model. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1597–1602, 2022.

[9] Lv. Chinese named entity recognition in the geoscience domain based on bert. *Earth and Space Science*, 2022.

[10] P. Patil, A. Ranade, M. Sabane, O. Litake, and R. Joshi. L3cube-mahaner: A marathi named entity recognition dataset and bert models. *arXiv preprint arXiv:2204.06029*, 2022.

[11] G. Puccetti, V. Giordano, I. Spada, F. Chiarello, and G. Fantoni. Technology identification from patent texts: A novel named entity recognition method. *Technological Forecasting and Social Change*, 186:122160, 2023.

[12] L. Tao, Z. Xie, D. Xu, K. Ma, Q. Qiu, S. Pan, and B. Huang. Geographic named entity recognition by employing natural language processing and an improved bert model. *ISPRS International Journal of Geo-Information*, 11 (12):598, 2022.

[13] A. Tikayat Ray, O. J. Pinon-Fischer, D. N. Mavris, R. T. White, and B. F. Cole. aerobert-ner: Named-entity recognition for aerospace requirements engineering using bert. In *AIAA SCITECH 2023 Forum*, page 2583, 2023.

[14] R. Yang, Y. Gan, and C. Zhang. Chinese named entity recognition based on bert and lightweight feature extraction model. *Information*, 13(11):515, 2022.