BDC:Using BERT and Deep Clustering to Improve Chinese Proper Noun Recognition

Yuanchi Ma
Beijing Institute of Technology
School of Computer Science and Technology
beijing,China
3220215164@bit.edu.cn

Hui He
Beijing Institute of Technology
Institute of Engineering Medicine
beijing,China
hehui617@bit.edu.cn

Zhendong Niu
Beijing Institute of Technology
School of Computer Science and Technology
beijing,China
zniu@bit.edu.cn

***Abstract*—Proper noun recognition is a sub-task in named entity recognition. However, few methods have been specifically applied to the Chinese. The reason is that most of the existing deep clustering methods rely on manually labeled training sets, which take a long time in the learning process. And due to the wide and large-scale nature of the proprietary domain and the lack of word boundaries, recognizing Chinese specialized terms from unstructured text remains challenging. In this paper, we design an unsupervised method to improve Chinese proper noun recognition. The first step is to implement the word separation for Chinese, followed by a BERT-based improved word characterization method to obtain word vectors. Finally, we use the autoencoder-based deep clustering method to complete the extraction of proper nouns from books. We have done comparison experiments on the public dataset and our selected professional book data respectively, and the result is an improvement of our method in both the accuracy and F1 values.** [1]

***Index Terms*—Proper noun recognition,BERT,Deep clustering,GMM**

I. Introduction

Proper noun recognition in NER is a key task that not only locates new terms in specialized fields but also identifies its entity classes from unstructured text, which can then be provided to various downstream NLP tasks for information acquisition. For example, question answering[1], relationship extraction [2], key information retrieval [3],[4], entity extraction and linking [5], hotspot discovery [6], etc. Compared with English, the task of extracting proper names for Chinese has been a major challenge due to the lack of word boundaries and large-scale manual annotation datasets. In detail, because of the unique language structure of the Chinese language, many word ambiguities occur [7], ignoring word-level information and using character-level information directly to recognize Chinese entities usually leads to poor performance, so Chinese word separation (CWS) needs to be performed to use word-level information to help determine word boundaries. For example, ”Agricultural Bank of China Xingtai Branch” is a proper noun in the banking world. However, we can also mark ”China”, ”Xingtai” and ”bank” as separate entities.

Many types of methods nowadays need a large amount of labeled data for training neural networks [8, 9], but for data in unknown specialized fields, we do not have labeled data available for training. For the above problems, this paper proposes an unsupervised method to improve Chinese proper noun recognition(BDC). The first initial division of Chinese proper nouns is achieved by performing Chinese word separation. And then the word vector is obtained in the word vector representation module using the improved word characterization method based on BERT [10]. Finally, the word characterization results are imported into the autoencoder-based deep clustering network module, and the key features are extracted and mapped to the two-dimensional vector representation space to realize the word clustering of all unknown nouns in Chinese data, thus completing the clustering and extraction of proper nouns in Chinese. The process is shown in Figure 1.

In summary, we have made the following contributions to this paper:

- To the best of our knowledge, this is the first work to improve the recognition of proper nouns in Chinese based on deep clustering networks.
- We transform the NER task into an unsupervised clustering task, which can directly identify and extract proper nouns without a large number of labeled training data.
- We deal with complex book data and provide a unique feature vector representation space for proper nouns. space.
- We have carried out extensive experiments on three open Chinese NER data sets and selected Chinese book data, and the results show that our model achieves better results than the baseline NER model.

In the rest of this article, we will introduce the relevant work, the relevant methods, and the role of our model in the second section. In the third section, we introduce the implementation technology and process of our method in detail. In the fourth section, the experimental process and results are introduced, and ablation experiments are conducted to study the importance of each part of BDC. Finally, the paper is summarized in the fifth section

II. Related Work

The proper noun extraction tasks under the NER task follow the NER task classification and can be mainly classified into supervised and unsupervised approaches.

[1] DOI No.10.18293/SEKE23-128

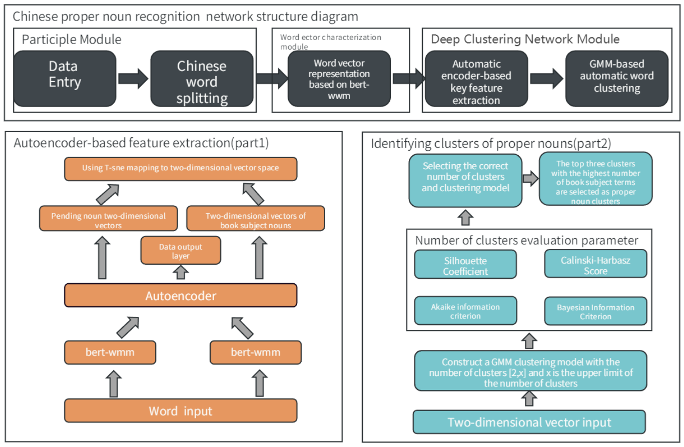

Chinese proper noun recognition network structure diagram
Participle Module
Data Entry
Chinese word splitting
Word ector characterization module
Word vector representation based on bert-wwm
Deep Clustering Network Module
Automatic encoder-based key feature extraction
GMM-based automatic word clustering
Autoencoder-based feature extraction(part1)
Using T-sne mapping to two-dimensional vector space
Pending noun two-dimensional vectors
Two-dimensional vectors of book subject nouns
Data output layer
Autoencoder
bert-wmm
bert-wmm
Word input
Identifying clusters of proper nouns(part2)
Selecting the correct number of clusters and clustering model
The top three clusters with the highest number of book subject terms are selected as proper noun clusters
Number of clusters evaluation parameter
Silhouette Coefficient
Calinski-Harbasz Score
Akaike information criterion
Bayesian Information Criterion
Construct a GMM clustering model with the number of clusters [2,x] and x is the upper limit of the number of clusters
Two-dimensional vector input

Fig. 1: This flowchart shows each of the three important parts of this algorithmic network.Participle Module is the first part, which segments the text part. word vector characterization module is the second part, which implements word segmentation and converts it into word vectors. deep Clustering Network Module is the third part, which is divided into two parts. They are autoencoder-based feature extraction (part1) and Identifying clusters of proper nouns (part2). Their roles are to extract word vector features and determine clusters of proper nouns after clustering, respectively.It should be noted that the adaptability of AE to bert has been proven to be better after extensive experiments.

A. *Supervised and unsupervised methods*

Among the supervised methods are dictionary-based methods, statistical machine learning-based methods, and deep learning-based methods. The unsupervised methods are mainly rule-based and clustering-based. Rule- and dictionary-based methods were the first methods with named entity libraries and manual rules. Unsupervised based methods such as [11], which is an unsupervised method based on cyclic consistency training. Statistical machine learning-based methods mainly include Hidden Markov Models [12], Maximum Entropy Models (MEM), Support Vector Machines [13], and Conditional Random Fields [14], where CRF is the most widely used statistical NER method.

Most deep learning-based methods [15] inherit the classical lstm-crf or cnn-crf architecture. [16, 17] use convolutional neural networks to learn word representations from English NER characters. One of them [16] proposed a semi-supervised learning model based on the BiLSTM neural network with a large amount of unlabeled text and a rather limited amount of labeled text. [18] proposed a gated convolutional neural network model for Chinese NER. [19] proposed a novel word character LSTM model for Chinese NER, which adds word information to the beginning or end characters of a word.

B. *proper noun recognition*

The above NER methods are very effective for generic NER vocabulary recognition, but not so effective for proper noun recognition in specialized fields. Recently, some professional domain proper noun recognition methods have been proposed, and they more or less borrow the ideas from the above methods. [20] A multilingual knowledge base based on Wikipedia is built to provide relevant contextual information for the Named Entity Recognition (NER) model to recognize proper nouns. To further improve the classification accuracy, word clustering information is added as feature embedding using K-means clustering method[21]. [22] Language model equipped with access to an external knowledge base (KB). Our Knowledge-Augmented Language Model continues this work by augmenting traditional models with KBs, which fall into the category of unsupervised. In contrast, our approach takes full advantage of the best-performing word characterization methods coming in and gets a mapping of their features to a two-dimensional space based on autoencoder, which yields more comprehensive results to improve the performance of proper noun recognition.

III. Method

A. *Participle Module*

We use jieba word division with the following principle: for existing words, a word graph scan is implemented based on a prefix lexicon to generate a directed acyclic graph (DAG) composed of all possible word formation cases of Chinese words in a sentence, and dynamic programming is used to find the maximum probability path to find the maximum cut-off combination based on word frequency. For unregistered words, an HMM model based on the word formation ability of Chinese characters is used, and the Viterbi algorithm is used for calculation.

B. *Siamese Network model based on BERT-wwm*

The BERT model [23] is one of the best-performing models in the field of word representation in recent years, and it can convert words into word vector representations, which is convenient for us to use in the next feature extraction work. [10] proposed a new improvement strategy based on the original BERT. We next used this model to achieve the characterization of all unknown nouns in the book data after using the overword for the text data extracted above. Their model proposes a new masking strategy, called MLM as a correction. In the original BERT, a word chunk tagger [10] is used to split the text into word chunk tags, where some words are split into several small fragments. Whole word masking mitigates the drawback of masking only a portion of the whole word, which is easier to predict for the model. In this model, we designed this part of the structure as the Siamese Network structure. As shown in Figure 1, part 1. That is, in training and testing, the sub-networks of the model use the BERT-wwm model, and the two models share parameters. The reason for this is to facilitate text standardization output and provide better vector space for the following in deep clustering.

C. Autoencoder-based feature extraction

For the obtained word vectors, we need to further obtain their key features. The role of the encoder [24] is to encode the high- dimensional input x into a low-dimensional hidden vari- able h. This forces the neural network to learn the most informative features. The role of the decoder is to reduce the hidden variable h in the hidden layer to its initial dimension. The best state is that the output of the decoder recovers the original input perfectly or approximately, i.e. $x^r = x$. Encoding process from the input layer to the hidden layer:

$$h = g_{\theta_1}(x) = \alpha(W_1 x + b_1) \tag{1}$$

The decoding process from the hidden layer to the input layer:

$$\hat{x} = g_{\theta_2}(h) = \alpha(W_2 x + b_2) \tag{2}$$

The optimization objective function of the algorithm is written as:

$$MinimizeLoss = dist(X, X^R) \tag{3}$$

where dist is the distance metric function of the two, which is usually calculated using the mean squared variance. If the number of neurons in the input layer n, is greater than the number of neurons in the hidden layer m, then we are equivalent to reducing the data from n to m di- mentions. Then we use this m-dimensional feature vector to reconstruct the original data.

But to visualize the data, we still need to reduce the di- dimensionality. The dimensionality reduction should preserve the local features of the data. Then the data is visualized using the T-SNE [25] method for dimensionality reduction. The algorithm defines a nonlinear mapping from the feature space M to the two-dimensional feature space Z. It minimizes the difference between the corresponding prob- ability distributions of the mismatch M and Z in terms of pairwise distances by minimizing the asymmetry. Thus the feature space Z obtained by these two consecutive steps is more suitable for estimating the number of clusters

D. Identifying clustering modules

We use a GMM-based clustering algorithm [26]. There are four main criteria for judging the number of clusters using this method: the BIC value [27] of the model, the AIC value [28], the silhouette coefficient [29] and the calinski harabasz value [30]. The first step is to create a GMM model with a pre-set range of k values for the number of clusters $k \subset [2, n]$, where $n = 3, 4, 5...$ After creating the set of models based on the range of K values, the BIC values and AIC values under each model are calculated as a direct means of model evaluation. Where the AIC value describes the appropriate number of clusters and the BIC value assesses the simplicity and validity of the model. Calculation of the silhouette coefficients and CH values is used as an indirect means of evaluating the model. Our goal is to find the AIC and BIC values of the highest point of the silhouette coefficient, the CH value, and the lowest point at the relevant k-value. However, these values are fluctuating because of the existence of local optima. Therefore, from a statistical point of view, a Monte Carlo method is introduced to take expectations to avoid local optima and to derive the optimal number of clusters. In this way, we can perform GMM clustering analysis on the data based on the optimal number of clusters obtained. Among them, the probability density function of the GMM model is as follows.

$$P_M(x) = \sum_{i=1}^{K} p(k)p(x \mid k) = \sum_{i=1}^{K} \alpha_k p(x \mid \mu_k, \sum\nolimits_k) \tag{4}$$

In this way, we can perform GMM clustering analysis on the data based on the optimal number of clusters obtained. In the GMM model of this paper, the aic value is calculated for each value of k. Let n be the number of observations and RSS be the residual sum of squares, then the AIC becomes as follows.

$$AIC = 2k + nln(RSS/n) \tag{5}$$

In BIC, k is the number of model parameters, n is the number of samples, and L is the likelihood function. $kln(n)$ penalty term is used in cases where the number of dimensions is too large and

$$BIC = kln(n) - 2ln(L) \tag{6}$$

In the silhouette coefficient, the intra-cluster dissimilarity is $a(i)$. The average of the dissimilarity of the i vector to other points in the same cluster reflects the cohesiveness. The inter-cluster dissimilarity is $b(i)$. The minimum of the average dissimilarity of the i vector to other clusters reflects the degree of separation.

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{7}$$

The CH values are calculated as follows. Where, n denotes the number of clusters,k denotes the current class, $trB(k)$ denotes the trace of the interclass departure matrix, and $trW(k)$ denotes the trace of the intra-class departure matrix.

$$CH(k) = \frac{trB(k)/(k-1)}{trW(k)/(n-k)} \tag{8}$$

At this point, we need to know the topic of the book has identified the clusters where the proper nouns exist. So next we perform a sub-word extraction and keyword extraction for the name of the book and the full text of the book respectively and select its extracted proper nouns and the first keyword as the subject words of the book. By mapping them to two-dimensional space by the above method, the category they belong to is the cluster of proper nouns we need to get at last.

IV. Experiments

A. datasets

Resume dataset. The dataset contains 4761 samples, 8 categories, for NAME, EDU, LOC, ORG, PRO, TITLE, CONT, and RACE.

CCKS2019 task1 dataset.Which has a total of 1379 samples, 6 categories, for anatomical sites, surgery, diseases and diagnoses, drugs, laboratory tests, and imaging tests

WeiBo NER dataset.The dataset contains 1890 samples with 7 categories as LOC.NAM, LOC.NOM, PER.NAM, ORG.NOM, ORG.NAM, GPE.NAM and PER.NOM.

200 book dataset. The 200 e-books collected by this study whose fields include computer science, chemistry, etc. Teaching books from various Chinese universities in specialized fields contain a large number of proper nouns that are well suited for this study.

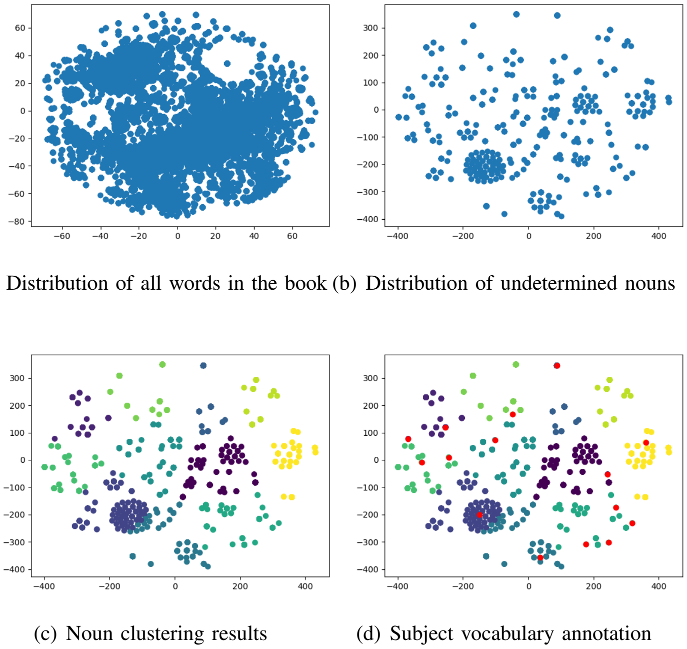
B. baseline

We compared our model on the three datasets mentioned above with the following open-source NER models. Since these baseline models were presented at different times, the different models apply to different datasets.

- [31]:Roberta-base-finetuned-cluener2020-Chinese
- [32]: Word2vec can express a word in vector form quickly and effectively through the optimized training model according to the given corpus
- [16]: It directly trains F1 scores instead of label accuracy, and trains F1 scores and label accuracy in a comprehensive way.
- [19]: It adds word information to the beginning or end character of a word and reduces the impact of word segmentation errors while obtaining word boundary information in the LSTM model of word character.
- [33]: The way to construct the graph neural network is to implement the Chinese NER as the graph node classification task through the dictionary.
- [34]: It proposes a simple and effective character-level Chinese NER representation method.
- [35]: This improves the performance of Chinese NER by integrating the structural information between Chinese characters.
- [36]: This is a method that combines Robert and CRF. Intended to help us explore detailed performance under different model structures.

C. Implementation details

In the experiments on the book dataset, to visualize the experimental results, we provide a demonstration of the clustering process of a book. As shown in Fig 2.Book is ”MySQL Database Design and Application”.The input layer is set to 768, and the hidden layer is set to 16 for the best results. The number of iterations is set to 100.



(a) Distribution of all words in the book (b) Distribution of undetermined nouns

(c) Noun clustering results (d) Subject vocabulary annotation

Fig. 2: Clustering results graph.

D. Experimental results

We tested three datasets on the baseline and BDC methods, respectively. Their accuracy, recall, and F1 values are calculated. From the experimental results in Tables 1 and 2, our method obtained the best F1 value on the CCKS2019 and books data sets respectively, especially on the book data set. The other two data sets have better baseline results. The reason we analyzed is that Resume and WEIBO NER data sets are more general and less professional, while CCKS2019 and books data sets in the computer field have more professional terms. They focus on the professional terms of the medical field and computer field respectively,

TABLE I: This table shows how BDC scored on two public datasets, People’s Daily 2014 and CCKS2019 task1. We calculated Precision, Recall, and F1 values separately.

baseline	Resume			CCKS2019 task1		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
[31]	93.02%	92.74%	92.88%	11.11%	58.85%	18.69%
word2vec[32]	-	-	85.89%	8.89%	**80.00**%	14.55%
WC-LSTM[19]	95.21%	95.10%	95.15%	51%	53%	55%
LGN[33]	95.20%	95.44%	95.36%	45%	43%	39%
LSTM[34]	95.44%	**95.76**%	95.60%	73%	71%	62%
MECT[35]	**96.45**%	95.38%	**95.91**%	36%	56%	30%
Roberta+CRF[36]	94.50%	95.00%	94.75%	31.32%	60.77%	41.34%
BDC	93.57%	84.72%	88.93%	**86.33**%	56.19%	**68.07**%

E. Ablation experiments

In the ablation experiment, we want to explore the effect of different modules on the overall network structure and results. In this algorithm, there are three modules. Among them, the

TABLE II: This table shows the scores of BDC on two hundred dedicated book datasets. We calculated Precision, Recall, and F1 values separately.

baseline	WEIBO NER			Book dataset		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
[31]	**72.38**%	**41.57**%	52.81%	42.23%	46.33%	44.19%
[16]	70.12%	40.66%	51.47%	36.13%	0.29%	42.05%
WC-LSTM[19]	66.10%	31.98%	43.11%	35.21%	34.88%	35.04%
LGN[33]	-	-	55.31%	48.67%	38.55%	43.02%
LSTM[34]	71.86%	49.11%	58.35%	57.32%	31.69%	40.81%
MECT[35]	-	-	61.91%	-	-	52.82%
Roberta+CRF[36]	70.44%	34.67%	46.47%	66.31%	50.11%	57.08%
BDC	72.17%	36.83%	48.77%	**75.65**%	**51.43**%	**60.29**%

participle module is a pre-processing of the book data, so we use the word vector representation module and the deep clustering network module as variables to explore which part has more influence on this algorithm.

We replace the word vector representation module and the deep clustering network module of the algorithm with other methods, respectively. The word vector characterization module is replaced by word2vec as the word characterization method, which we call BDC-2vec. The deep clustering network module uses k-means instead of GMM clustering, and instead of using the method of automatically determining the number of clusters based on parameters, only the contour coefficients are used as the criterion for determining the number of clusters. We call it BDC-kmeans.

According to Table 3, we can see that BDC-2vec works better than BDC-kmeans and the original BDC works last. Thus for the overall algorithm, the deep clustering network module has more impact and this is where we innovate the most.

V. Conclusion

In this paper, we propose a large-scale network architecture for the Chinese language to extract the technical terms among them. We combine a pre-training model based on BERT improvement with an autoencoder based deep clustering model to achieve the transition from a proper noun extraction task to a clustering task. We design multiple working modules to extract key features from the learned information for proper noun prediction. Extensive experiments show that our proposed method reaches at least the state-of-the-art baseline. denotes the effectiveness in a good representation of this unsupervised proper noun extraction method. In addition, the word vector representation of the pre-trained model is slow due to the large vocabulary of many books' data. In the future, we will investigate how to reduce the model parameters and improve the training speed while enhancing the experimental results.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62272048.

References

[1] A. Fader, L. Zettlemoyer, and O. Etzioni, "Paraphrase-driven learning for open question answering," in *ACL (1)*. The Association for Computer Linguistics, 2013, pp. 1608–1618.

[2] W. U. Ahmad, N. Peng, and K. Chang, "GATE: graph attention transformer encoder for cross-lingual relation and event extraction," in *AAAI*. AAAI Press, 2021, pp. 12 462–12 470.

[3] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *ACL (1)*. The Association for Computer Linguistics, 2015, pp. 167–176.

[4] Y. Zhang, H. Fei, and P. Li, "End-to-end distantly supervised information extraction with retrieval augmentation," in *SIGIR*. ACM, 2022, pp. 2449–2455.

[5] Z. Zhang, X. Sind, T. Liu, Z. Fang, and Q. Li, "Joint entity linking and relation extraction with neural networks for knowledge base population," in *IJCNN*. IEEE, 2020, pp. 1–8.

[6] J. Jia, S. Tan, G. Ji, and B. Zhao, "Discovering hotspots in dynamic spatial networks using mobility data," in *CBD*. IEEE, 2019, pp. 102–107.

[7] P. Zhu, D. Cheng, F. Yang, Y. Luo, D. Huang, W. Qian, and A. Zhou, "Improving chinese named entity recognition by large-scale syntactic dependency graph," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 979–991, 2022.

[8] X. Ma and E. H. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," in *ACL (1)*. The Association for Computer Linguistics, 2016.

[9] J. Li, B. Chiu, S. Feng, and H. Wang, "Few-shot named entity recognition via meta-learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4245–4256, 2022.

[10] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese BERT," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3504–3514, 2021.

[11] A. Iovine, A. Fang, B. Fetahu, O. Rokhlenko, and S. Malmasi, "Cyclener: An unsupervised training approach for named entity recognition," in *WWW*. ACM, 2022, pp. 2916–2924.

[12] R. S. D. M. Bikel an'dS. Miller and R.Weischedel, "Nymble:ahigh- performance learning name-finder," in *Proc. 5th Conf. Appl.* Natural Lang. Process, 1998, p. 194201.

[13] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *COLING*, 2002.

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*. Morgan Kaufmann, 2001, pp. 282–289.

[15] F. Ullah, I. Ullah, and O. Kolesnikova, "Urdu named entity recognition with attention bi-lstm-crf model," in

TABLE III: This table shows the results of the ablation experiment. Our experimental design references [37]. BDC-2vec indicates that the feature extraction neural network is replaced. BDC-kmeans indicate that the clustering module is replaced.

PARTS		People's Daily 2014			200 books			800 books		
BDC-2vec	BDC-kmeans	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
✗	✓	95.15%	95.21%	51%	36%	58%	34.12%	32.00%	56.00%	40.73%
✓	✗	95.39%	95.98%	36%	57.32%	31.69%	51.32%	58.33%	33.09%	42.22%
✓	✓	73.3%	41.57%	60.32%	52.23%	56.33%	35.12	50.87%	35.10%	41.54%%

MICAI (2), ser. Lecture Notes in Computer Science, vol. 13613. Springer, 2022, pp. 3–17.

[16] H. He and X. Sun, "F-score driven max margin neural network for named entity recognition in chinese social media," in *EACL (2)*. Association for Computational Linguistics, 2017, pp. 713–718.

[17] Z. Zhai, D. Q. Nguyen, and K. Verspoor, "Comparing CNN and LSTM character-level embeddings in bilstm-crf models for chemical and disease named entity recognition," in *Louhi@EMNLP*. Association for Computational Linguistics, 2018, pp. 38–43.

[18] C. Wang, W. Chen, and B. Xu, "Named entity recognition with gated convolutional neural networks," in *CCL*, ser. Lecture Notes in Computer Science, vol. 10565. Springer, 2017, pp. 110–121.

[19] W. Liu, T. Xu, Q. Xu, J. Song, and Y. Zu, "An encoding strategy based word-character LSTM for chinese NER," in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 2379–2389.

[20] X. Wang, Y. Shen, J. Cai, T. Wang, X. Wang, P. Xie, F. Huang, W. Lu, Y. Zhuang, K. Tu, W. Lu, and Y. Jiang, "DAMO-NLP at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition," in *SemEval@NAACL*. Association for Computational Linguistics, 2022, pp. 1457–1468.

[21] J. Laishram, K. Nongmeikapam, and S. Naskar, "Deep neural model for manipuri multiword named entity recognition with unsupervised cluster feature," in *ICON*. NLP Association of India (NLPAI), 2020, pp. 420–429.

[22] A. Liu, J. Du, and V. Stoyanov, "Knowledge-augmented language model and its application to unsupervised named-entity recognition," *CoRR*, vol. abs/1904.04458, 2019.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[24] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *CVPR*. IEEE Computer Society, 2017, pp. 2770–2779.

[25] J. Cheng, H. Liu, F. Wang, H. Li, and C. Zhu, "Silhouette analysis for human action recognition based on supervised temporal t-sne and incremental learning," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3203–3217, 2015.

[26] S. Messaoud, A. Bradai, and E. Moulay, "Online GMM clustering and mini-batch gradient descent based optimization for industrial iot 4.0," *IEEE Trans. Ind. Informatics*, vol. 16, no. 2, pp. 1427–1435, 2020.

[27] C. A. C. Lozada, C. Montealegre, M. Mejia, M. Mendoza, and E. León, "Web document clustering based on a new niching memetic algorithm, term-document matrix and bayesian information criterion," in *IEEE Congress on Evolutionary Computation*. IEEE, 2010, pp. 1–8.

[28] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[30] S. Lukasik, P. A. Kowalski, M. Charytanowicz, and P. Kulczycki, "Clustering using flower pollination algorithm and calinski-harabasz index," in *CEC*. IEEE, 2016, pp. 2724–2728.

[31] L. D. Thomas Wolf and V. Sanh.et.al, "Transformers: State-of-the-art natural language processing." Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[32] C. G. e. a. Mikolov T, Chen K, "Efficient estimation of word representations in vector space." arXiv preprint, 2013, pp. arXiv:1301.3781, 2013.

[33] T. Gui, Y. Zou, and Q. Z. .et.al, "A lexicon-based graph neural network for chinese NER," in *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 2019, pp. 1040–1050.

[34] R. Ma, M. Peng, Q. Zhang, Z. Wei, and X. Huang, "Simplify the usage of lexicon in chinese NER," in *ACL*. Association for Computational Linguistics, 2020, pp. 5951–5960.

[35] S. Wu, X. Song, and Z. Feng, "MECT: multi-metadata embedding based cross-transformer for chinese named entity recognition," in *ACL/IJCNLP (1)*. Association for Computational Linguistics, 2021, pp. 1529–1539.

[36] A. Chernyavskiy, D. Ilvovsky, and P. Nakov, "aschern at semeval-2020 task 11: It takes three to tango: Roberta, crf, and transfer learning," *CoRR*, vol. abs/2008.02837, 2020.

[37] B. Shi, L. Ji, P. Lu, Z. Niu, and N. Duan, "Knowledge aware semantic concept expansion for image-text matching," in *IJCAI*. ijcai.org, 2019, pp. 5182–5189.