Implementation of Chinese Named Entity Recognition Method Based on Bi-LSTM and CRF

Authors: Jiao Xinyu

Date: 2024-01-03T00:00:00+00:00

Abstract

Chinese Named Entity Recognition (NER) plays a crucial role in the field of natural language processing and is essential for improving the effectiveness of information extraction and text understanding. This paper implements a Chinese named entity recognition method based on Bidirectional Long Short-Term Memory (Bi-LSTM) networks and Conditional Random Fields (CRF). First, we extract features from the input Chinese text through a Bi-LSTM network, leveraging the advantages of bidirectional sequence learning to capture contextual information, thereby achieving a better understanding of context. This network can effectively handle long-distance dependencies, enhancing the ability to recognize complex structures and nested entities in text. Second, CRF is introduced as the decoding layer for the sequence labeling task to perform global optimization on the Bi-LSTM output, taking into account the relationships between entity labels. The CRF model captures global dependencies in label sequences, which helps correct local errors and improves the overall performance of the NER system. Experimental results demonstrate that the Bi-LSTM-CRF model exhibits excellent performance on Chinese NER tasks, achieving favorable results across various metrics.

Full Text

Preamble

Implementation of Chinese Named Entity Recognition Method Based on Bi-LSTM and CRF

Name: Jiao Xinyu
Class: Bio-Research 2302
Student ID: 2023201249

1.1 Research Background and Significance

Named Entity Recognition (NER) is one of the fundamental tasks common to many high-level Natural Language Processing (NLP) applications. Its primary objective is to identify and classify entities with specific meanings from text, typically including person names, organizations, locations, dates, times, currencies, percentages, and other concrete named entities. NER plays a crucial role in numerous NLP technologies such as information retrieval, machine translation, and question-answering systems. To address these challenges, researchers and practitioners have employed various techniques, including rule-based methods, machine learning approaches (such as Conditional Random Fields and sequence labeling models), and the recently popular deep learning methods.

1.2 Related Research Status

In early NER research, due to the limited scale of available data, the predominant approach was rule-based systems developed by linguistic experts. These systems relied on manually crafted rules derived from detailed analyses of the grammatical and morphological structures, contextual collocations, and lexical patterns of named entities. During the middle research period, probabilistic and statistical methods became mainstream. Researchers designed models using statistical and probabilistic knowledge for specific tasks, performed supervised learning from large amounts of annotated data to obtain trained models, and then used these models to identify named entities in sentences. Commonly used statistical models included Hidden Markov Models, Conditional Random Fields, and Support Vector Machines. In recent years, deep learning-based NER methods have achieved performance comparable to statistical models, and in most cases have surpassed them, by constructing artificial neural networks to model nonlinear processes and automatically learn abstract features from different datasets. Moreover, deep learning methods exhibit lower dependence on manual features and stronger cross-domain adaptability.

1.3 Research Content of This Paper

The specific objective of Named Entity Recognition is to identify and classify entities with particular meanings from given text. These entities can be any named entities in the text, typically including the following categories: (1) Person names, such as “Jiang Xiaobai”; (2) Organizations, such as “People’s Congress”; (3) Locations, such as “Beijing”; and (4) Dates, such as “2023”. The goal of NER is not only to identify these entities but also to classify them into the appropriate categories. Therefore, NER output typically consists of a series of entity tags, each corresponding to an entity in the text and indicating its type. In this study, we primarily aim to gain in-depth understanding of the performance characteristics of deep learning-based NER models across different dimensions, thereby providing valuable reference for model selection in practical applications.

2.1 Conditional Random Fields

The BIO tagging scheme is commonly adopted in Chinese named entity recognition tasks. NER can be treated as a sequence labeling problem, for which linear-chain Conditional Random Fields are frequently employed. For a sentence sequence X = (x_1, x_2, ..., x) consisting of n characters and a label sequence Y = (y_1, y_2, ..., y), each character's label y can be selected from a predetermined set of tags. When each character x is assigned a label y , Y forms a random field. Furthermore, when a constraint is added to the label selection process such that tag y depends only on its previous tag y $_1$, this random field is called a Markov random field. For the two parameters in the Markov random field—the input sequence X and label sequence Y—where X is the given text sequence and Y is the sequence labeling output conditioned on X, P(Y|X) constitutes a Conditional Random Field. CRF requires defining appropriate feature functions (typically real-valued functions) to represent both transition and state features.

2.2 Long Short-Term Memory Networks

The LSTM network is an improved architecture based on RNN that incorporates various gated memory units within its neural cells to control current input and previous output. During training, an LSTM cell can regulate the previous moment's output, current moment's input, and the state up to the current moment through its forget gate, output gate, and input gate. Through these three gating units, LSTM networks acquire selective memory capabilities.

2.3 Bi-LSTM-CRF Model

The Bi-LSTM-CRF model is a deep learning-based named entity recognition model that combines Conditional Random Fields (CRF) with Bidirectional Long Short-Term Memory networks (Bi-LSTM). The model architecture consists of three layers: an embedding layer, a Bi-LSTM layer, and a CRF layer, as illustrated in [Figure 1: see original paper].

Embedding Layer: This "vectorization" layer primarily transforms high-dimensional sparse vectors into dense vectors to facilitate downstream processing.

Bi-LSTM Layer: LSTM is a variant of RNN whose design makes it particularly suitable for modeling sequential data such as text. Bi-LSTM combines forward and backward LSTM networks, both of which are commonly used in natural language processing tasks to model contextual information. While using LSTM to model sentences presents the problem of being unable to encode information from back to front, Bi-LSTM can better capture bidirectional semantic dependencies.

CRF Layer: The main function of this layer is to constrain the output from the Bi-LSTM layer to avoid unreasonable label sequences. After the Bi-LSTM network generates probability transition vectors, these vectors and labels form

a conditional random field. Once the model is trained to minimize the objective loss function, the CRF layer employs the Viterbi algorithm—a dynamic programming algorithm—to find the most reasonable sequence from the entire conditional transition matrix as the predicted labels.

Optimization Algorithm: During neural network training, a common phenomenon occurs where the model achieves high accuracy on training data but significantly lower accuracy on test data. This situation is called overfitting, which becomes more pronounced when training data is insufficient while network parameters are numerous. An overfitted model is generally considered to have no reference value, so training must avoid producing such models. Dropout is used to address overfitting by randomly deactivating neurons with a certain probability during each training iteration. The working principle of dropout is shown in [Figure 2: see original paper].

3. Implementing Chinese NER with Bi-LSTM-CRF Model

3.1.1 Dataset

The MSRA Chinese news dataset, released by Microsoft Research Asia for Chinese named entity recognition, primarily includes location names, organization names, and person names using the BIO tagging strategy. The training set contains 42,000 sentences with over 3.4k location entities (LOC), 1.9k organization entities (ORG), and 1.6k person entities (PER). The validation and test sets are each approximately one-tenth the size of the training set.

3.1 Data Processing Pipeline

1. **Corpus and Label Reading:** Read the corpus and label information to establish mapping dictionaries for vocabulary and tags. The following is an example of label mapping: {'O': 0, 'B-ORG': 1, 'I-PER': 2, 'B-PER': 3, 'I-LOC': 4, 'I-ORG': 5, 'B-LOC': 6}.
2. **Dataset Loading:** Load the training, validation, and test sets separately. The BertTokenizer is used to convert raw text into encodings. This process does not involve converting words into word vectors but merely tokenizes the raw text, adds special tokens such as [MASK], [SEP], and [CLS], and then converts these tokens into dictionary indices.
3. **Data Structure:** The processed data consists of three components: **input_{ids}** (a matrix of vocabulary encodings padded with 0), **label_{ids}** (a matrix of corresponding label encodings padded with 0), and **mask** (positions of valid tokens in the input sequence marked as 1, others as 0).

3.2 Evaluation Methods

Precision: Also called positive predictive value, representing the proportion of actual positive samples among those predicted as positive.
Recall: Also called sensitivity, representing the proportion of actual positive samples correctly predicted as positive among all positive samples in the dataset.
F1_{score}: The harmonic mean of precision and recall.
Accuracy: The proportion of correctly classified samples among the total sample size.
Multi-class Metrics (macro, micro, weighted): In multi-class tasks, each class has its own precision, recall, and F1 score, requiring balancing across classes to obtain global evaluation metrics:
- **Macro:** Calculate metrics for each class individually and then average them, ignoring class imbalance.
- **Micro:** Treat all samples as an aggregate, with denominators being the total number of samples.
- **Weighted:** Use the frequency of each class in the ground truth as weights.

3.3 Experimental Parameter Settings

Training follows a four-step strategy, with step 5 implementing an alternating training and testing approach. The parameter configuration is as follows: - embed_{size}=100: Embedding layer dimension - hidden_{size}=128: Hidden layer dimension - num_{layer}=1: Number of LSTM layers - lr=0.001: Learning rate - optimizer=Adam: Optimization algorithm - loss=CrossEntropyLoss: Cross-entropy loss function

3.4.1 Learning Rate

A learning rate of 0.001 produces smoother curve variations and yields better model performance.

3.4.2 Dropout Value

Initial experiments exhibited overfitting, prompting experiments with different dropout values. As shown in [Figure 3: see original paper] (loss curve without dropout) and [Figure 4: see original paper] (loss curve with dropout=0.5), using dropout demonstrates superior performance compared to no dropout. This study adopts dropout=0.5 for all experiments. Although dropout mitigates overfitting, the model still exhibits some overfitting behavior. Consequently, L1 regularization was experimented with, but despite improving model fit, it significantly reduced prediction precision and other metrics, leading to its abandonment.

3.4.3 Other Metrics

[Figure 5: see original paper] shows the loss curve after applying L1 regularization, [Figure 6: see original paper] displays the loss curve for 20 training epochs, and [Figure 7: see original paper] presents precision, recall, F1 score, and other evaluation metrics.

3.4.4 Experimental Results Analysis

[Figure 8: see original paper] provides prediction examples. The evaluation metrics analysis reveals low recognition accuracy for organizations (ORG), likely due to the high proportion of compound words where organization names contain person or location names, causing recognition errors. For instance, "美中关系全国委员会" is incorrectly segmented into "美中" and "全国委员会". Examination of CRF state transition features confirms the existence of I-LOC→B-ORG and I-PER→B-ORG transitions, indicating that the model has learned these patterns from the dataset and validating that organization names containing person and location names indeed cause recognition errors, as shown in [Figure 9: see original paper].

The observed overfitting and relatively low accuracy may stem from the small dataset size, which is somewhat insufficient for deep learning models that typically require large samples. From an application perspective, if the recognition task does not heavily depend on long-term contextual information, RNN-based models may introduce unnecessary complexity.

Conclusion

With the rapid development of communication technology and the internet, the scale of online data continues to expand. While this data contains substantial usable information, its unstructured format makes direct utilization difficult. As an important research field in text signal processing, information extraction aims to automatically extract structured information from unstructured data through computational methods. Named entities constitute a crucial component of this structured information, and accurate recognition directly impacts the effectiveness of information extraction technologies. Furthermore, research on Chinese corpus NER has primarily been concentrated domestically, with a later start and relatively outdated techniques compared to some developed countries, making research on named entity recognition—especially Chinese NER—extremely valuable and significant.

This study employs a deep learning-based Bi-LSTM-CRF model for Chinese named entity recognition, achieving a final F-score of 84%. However, due to dataset limitations, several directions warrant further investigation. Future optimization should focus on expanding the dataset, adjusting the model based on actual data characteristics, or optimizing the model architecture itself.

Note: Figure translations are in progress. See original paper for figures.