Research on Express Information Extraction Based on Multiple Sequence Labeling Models

Bihui Yu¹, ² and Ke Wang¹, ², *

¹Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang, China
²University of Chinese Academy of Sciences, Beijing, China

*Corresponding author: wonkers@sict.ac.cn

Abstract. With the popularity of online shopping, the logistics industry has developed rapidly. Major logistics companies have greater business needs to improve the efficiency of express sorting centers. Therefore, the technology of optimizing the efficiency of the express sorting system in the logistics industry and reducing the cost of time for customers to fill out express orders has received increasing attention from the society. The research work carried out in this paper is based on deep learning using natural language processing technology to extract express single text information, structured extraction of names and mobile phone numbers and addresses. The main content is that the express information structured extraction system automatically extracts information such as name, telephone, province, city, district, and detailed address. The docking of the express delivery sorting business system will help improve the logistics industry's operational capabilities. In this paper, four sets of sequence labeling models are constructed. Through multiple sets of data evaluation, it is shown that the model based on ERNIE and the learning rate decay strategy using Adam optimizer is superior to the current CNN, RNN, LSTM models. The highest F1 of the model constructed in this paper reaches 0.99473.

Keywords: Information extraction, Sequence annotation, Bert, Ernie.

1. Introduction

Online Shopping is popular now. Thus, logistics business is developing rapidly. Those logistics companies have greater business needs to improve the efficiency of express sorting centers. On the other hand, Consumers also hope that they can fill in their personal shopping receipt information more freely when adding a delivery address on the online shopping platform.

In order to fulfill this demand, the express information structured extraction system designed in this paper automatically extracts names, telephones, provinces, cities, districts, detailed addresses, etc., to improve the comfort of customers to fill in the information. In this paper, the express document information corpus labeled based on the BIO sequence labeling system is first used to represent and learn using word vectors (word embedding), using a recurrent neural network, long and short-term memory network.
The test results are evaluated in the test set, and the predicted results are evaluated using Precision, Recall and F1. Finally, based on the transfer learning idea, using Google BERT and Baidu BRNIE two pre-training models with good effect extraction, combined with the optimization of the learning rate attenuation strategy based on Adam optimizer, and finally built the express single information extraction model F1 Reached 0.99473.

2. Related Work

2.1. Information Extraction

Information Extraction is a processing technology that extracts information such as entities, relationships, and events from the text of daily communication in human society. Information extraction tasks are generally divided into four subtasks: named entity recognition, entity disambiguation, relationship extraction, and event extraction.

The MUC meeting divided the specific specifications of the information extraction task into four sub-tasks: named entity recognition task (NE), entity relationship extraction task (TR), argument role recognition (ST), template element filling (TR), determining the referential relationship [1].

The sub-tasks specified by ACE2005 for event extraction include: 1. Entity Detection and Recognition (EDR). The specific task is to identify entities and entity attributes. The attributes of the entity include category, type, and subtype. 2. Relationship Detection and Recognition (RDR), the specific task is to detect and identify the relationship and relationship attributes between two entities and identification arguments. The attributes of the entity relationship include tense, modality, type, and subtype. 3. Event detection and recognition (VDR), the specific task is to detect and identify specific types of events, argument roles, and event attributes. The attributes of the event include tense, modality, type, subtype, universality and tendency. 4. Value detection and identification. The specific task is to identify and normalize specific types of values. Such as time expressions such as age, phone number, time point, time period, etc.

2.2. Named Entity Recognition

Name Entity Recognition is the basic work of information extraction task. [2] Named entities are entities recognized by computers that can express the specific semantic information that others can receive in the text of daily communication in human society. The corresponding form of linguistics is nouns. The tasks of named entities specified by the MUC meeting include: Numeric expressions such as person names, organization names, place names, dates, times, monetary numeric percentage values, and other proper nouns.

The early recognition of named entities was influenced by two major schools of empiricism and empiricism [3]. The research directions of academia are basically divided into: relying on rules and relying on statistics. Relying on the statistical method is to use statistical methods and probabilistic theory to build a complex mathematical structure model for a specific named entity recognition task, and use the learning features of the labeled corpus and the parameters of the training model to train the statistical model. Recognize named entities in unlabeled corpora.

The method of extracting named entities that relies on rules is to use the existing symbol processing system and rule framework to construct a program that combines lexical syntax and semantic rules. The program analyzes the input sentences according to the specified rules, infers the phrases that may be named entities according to the manually constructed finite state machine, and then uses the pattern matching scheme to classify the entities [4].

3. Express Information Extraction System

We first build three classical models of recurrent neural network RNN, long and short-term memory network LSTM to predict and evaluate the effect, then try to use ERNIE and BERT pre-trained models for transfer learning, and use a variety of loss functions and the optimizer fine-tunes the model to optimize the construction of a custom model for extracting express delivery information in the e-
commerce logistics field. The following will introduce the four model structures and the corresponding effects.

3.1. Recurrent Neural Network
RNN (Recurrent Neural Network) is a neural network that is often used to process sequence data. In natural language processing, RNN can "remember" the historical information input by the sequence, so that it can better model the entire sequence semantically.

![Figure 1. The model architecture of RNN.](image)

The structure of the RNN model is also divided into input layer (Input Layer), hidden layer (Hidden Layer) and output layer (Output Layer). It should be noted that there is a reflowing arrow in the hidden layer. It is the role of this arrow that makes the RNN have the ability to "remember". Each arrow represents a transformation, which means that the arrows are connected with weights, and the neurons in the hidden layer are also weighted. In other words, as the sequence continues to advance, the hidden layer in the front will affect the hidden layer in the back. Obviously, "loss" is also constantly accumulating.

The calculation method of recurrent neural network is as follows:

\[
O_t = g(V \cdot S_t)
\]

\[
S_t = f(U \cdot X_t + W \cdot S_{t-1})
\]

The RNN model is used to predict the data set of the unknown label, and the prediction results of the model are shown as follows:

- Yu Xiaogang/P Yunnan Province/A1 Chuxiong Yi Autonomous Prefecture/A2 Nanhua County/A3 No. 37 Gucheng Road, East Street/A4 18513386163/T 1342638135/T
- Kou Mingzhe/P Heilongjiang Province/A1 Qitaihe City/A2 Taoshan District/A3 Fengcai Road Chaoyang Square/A4

In this paper, 1601 pieces of data are used, with batch_size=100 and epoch=30, and 480 steps of training. The final test results of the RNN model are P: 0.85951, R: 0.89996, and F1: 0.87927.

3.2. Long Short Term Memory
LSTM (Long Short Term Memory) is a special RNN structure that can learn long-term dependencies, it effectively solves the problem of gradient disappearance and gradient explosion during long sequence training. Compared to ordinary RNN, LSTM can perform better in longer sequences.
The key to LSTM is the Cell State. Their linear interaction is minimal, and it runs through the whole process. This mechanism enables the information transmitted in the LSTM structure to be smoothly transmitted on the entire link without sending changes.

Using the LSTM model to make predictions, the obtained model prediction results are shown as follows:

- Taiwan/A1 Chiayi County/A2 Fanlu Township/A3 Gonglu Village, Fanlu Township/A4 17zhi/T 19/T Xuan Shuyi/P 13720072123/T
- Ao Daqin/P Shanxi Province/A1 Linfen City/A2 Xi County/A3 South Street 18500509799/A4

In this paper, 1601 pieces of data are used, with batch_size=100 and epoch=30, and 480 steps of training.

The final test results of the LSTM model are P: 0.87430, R: 0.91215, and F1: 0.89282.

3.3. **BERT**

The language representation model BERT [5], which represents the bidirectional encoder representation of the converter. BERT pre-trains deep bidirectional representations through joint adjustments in the context of all layers.

BERT solves the aforementioned one-way constraint by proposing a new pre-training target: the "masking language model" (MLM), inspired by the gestalt task (Taylor, 1953). The masked language model randomly masks some tags from the input, and the goal is to predict the original vocabulary id of the masked word based only on its context.

The forecast results section shows:

- 19880996524/T Ge Cheng/P Chongqing City/A1 Zhongxian County/A3 No.13 Letianzhi Road/A4

In this paper, the final test result of the express delivery information extraction system built by the BERT model is: F1=0.98712 precision=0.98481 recall=0.98946.

3.4. **ERNIE**

The ERNIE model [6] itself maintains modeling based on word feature input, so that the model does not need to rely on other information when applied, and its model is more versatile and extensible.
Relative to the word feature input model, word features can model the combined semantics of words. For example, when modeling red, green, and blue words that represent colors, the semantic relationship between words can be learned through the semantic combination of the same word.

In addition, ERNIE’s training corpus introduces knowledge of multi-source data. In addition to the modeling of encyclopedia articles, it also learns about news information and forum dialogue data. The learning of dialogue data is an important way of semantic representation, and the Query semantics corresponding to the same reply are often similar.

The forecast results section shows:

- Heilongjiang Province/A1 Shuangyashan City/A2 Jianshan District/A3 40 meters north of the intersection of Bama Road and East Parallel Road/A4 Wei Yetao/P 18600009172/T

In this paper, the final test result of the express delivery information extraction system built by ERNIE model is: $F_1=0.99473$ precision=$0.99354$ recall=$0.99593$.

4. Experiments

4.1. Data sets

Because the courier note in the field of e-commerce logistics studied in this paper involves the personal privacy of name and phone address, the idea of randomly generating a splicing structure is adopted. The corpus generation rules are in line with reality, and the three types of information such as name, phone and address are randomly spliced during splicing. The data set is divided into training set train, verification set dev, and test set test.

4.2. Preprocessing process

First extract sentences and tags from the original data file to construct sentence sequences and tag sequences. In this paper, the entities that can be extracted from the express delivery information in the field of e-commerce logistics are labeled according to the name (person, P label), telephone (telephone, T label), address (address, A label) combined with the BIO system. Then convert the special characters in the sentence sequence. Obtain the integer index corresponding to the word according to the dictionary.

The data set category set of the express delivery information extraction system built in this article is:

```
{"B-P": 0, "I-P": 1, "B-T": 2, "I-T": 3, "B-A1": 4, "I-A1": 5, "B-A2": 6, "I-A2": 7, "B-A3": 8, "I-A3": 9, "B-A4": 10, "I-A4": 11, "O": 12}
```
Table 1. Label definition of express information extraction.

| LABEL | DEFINITION                  |
|-------|-----------------------------|
| P-B   | NAME START POSITION         |
| P-I   | MIDDLE OR END POSITION OF NAME |
| T-B   | PHONE START POSITION        |
| T-I   | MIDDLE OR END POSITION OF PHONE |
| A1-B  | PROVINCE START POSITION     |
| A1-I  | MIDDLE OR END POSITION OF PROVINCE |
| A2-B  | CITY START POSITION         |
| A2-I  | MIDDLE OR END POSITION OF CITY |
| A3-B  | COUNTY START POSITION       |
| A3-I  | MIDDLE OR END POSITION OF COUNTY |
| A4-B  | ADDRESS START POSITION      |
| A4-I  | MIDDLE OR END POSITION OF ADDRESS |
| O     | UNCONCERNED WORDS           |

4.3. Evaluation index

For the prediction results of each sequence sample, the sequence labeling task combines and evaluates the prediction results according to chunks. Evaluation indicators are usually Precision, Recall and F1.

- Precision. It is obtained by dividing the number of correct predictions of the model by the total number of predictions of the model. Pay attention to whether the results predicted by the model are accurate.
- Recall. It is obtained by dividing the number of correct predictions of the model by the number of real labels, paying attention to what the model missed.
- F1. It is a comprehensive evaluation index, considering both Precision and Recall, which is a compromise between Precision and Recall.

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]  

4.4. Build pre-trained models and fine-tune

Based on the idea of transfer learning, using the pre-trained model with excellent feature extraction work to learn the semantic information in the massive data to assist the express information structure extraction task of the express document data set in this paper. This paper attempts to use ERNIE and BERT to optimize the training model, and adopts Adam weight attenuation strategy to optimize the migration task to avoid overfitting the model. After fine-tuning into a custom model for structured extraction of express information in the field of e-commerce logistics, the effectiveness of the model is evaluated.

5. Results

Table 2. Comparison of the effects of express information extraction model.

| Model | Precision | Recall  | F1 Score |
|-------|-----------|---------|----------|
| RNN   | 0.85951   | 0.89996 | 0.87927  |
| LSTM  | 0.87430   | 0.91215 | 0.89282  |
| BERT  | 0.98481   | 0.98946 | 0.98712  |
| ERNIE | 0.99354   | 0.99593 | 0.99473  |
6. Conclusion
This paper first studies the three classic models of RNN, long-short-term memory network LSTM models to predict and evaluate the effect, then attempts to use ERNIE and BERT pre-trained models for transfer learning, and adopts Adam weight attenuation strategy to fine-tune the model optimize and build a custom model for express information extraction in the field of e-commerce logistics. Future work may consider introducing a multi-head attention mechanism to the Bi-LSTM-CRF model, trying to use Adamw and Amsgrad strategies for model optimization, and for pre-trained model ideas, you can try to use knowledge distillation to optimize transfer learning performance.

References
[1] A. Lavelli, M. E. Califf, F. Ciravegna et al. IE evaluation: Criticisms and recommendations. proceedings of AAAI2004 on the Workshop of Adaptive Text Extraction and Mining (ATEM), San Jose, U.S.A, 2004: 279-299.
[2] Andrew Mc Callum S. Sekine. Named Entity: History and Future. Technical report, 2004.
[3] Church K W, Mercer R L. Introduction to the Special Issue on Computational Linguistics Using Large Corpora [J]. Computational Linguistics, 1993, 19 (1): 1-24.
[4] Humphreys K, Gaizauskas R, Azzam S, et al. Description of the LaS IE-II System as Used for MUC-7 [C]. In Proceedings of the 7th Message Understanding Conference (MUC-7), 1998.
[5] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. 2018.
[6] Sun Y, Wang S, Li Y, et al. ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [J]. 2019.