Research on Text Mining of Railway Safety Supervisors Performance Based on BiLSTM and CRF

Fan Gao1,a  Shisheng Wang2,b  Xinqin Li3,c  Hongfei Cao4,d  Xiaolei Cai5,e

1China Academy of Railway Sciences
Beijing
China
86+18810137878

2China Academy of Railway Sciences
Beijing
China
86+13671068976

3China Academy of Railway Sciences
Beijing
China
86+13691460413

4China Academy of Railway Sciences
Beijing
China
86+18611908709

5China Academy of Railway Sciences
Beijing
China
86+13683667375

a1ight870202@163.com  bmrwsh@sina.com  ctyshi@rails.cn
dcaohongfeidut@163.com  ebaodao_1@sina.com

ABSTRACT: In the era of big data with unstructured data, the potential value of unstructured data mining is an effective way to solve the industry problem. Based on a large number of data recorded from the actual text of the performance plan of railway security supervisors, this paper proposes a text similarity calculation method based on text mining technology to calculate whether the personnel performance plan matches the reality. First, the text Named Entity Recognition (NER) is realized by using the combination of Bi-directional Long Short Term Memory (BiLSTM) and the Conditional Random Fields (CRF). Then based on the concept similarity calculation of HowNet, we designed the method of calculating the similarity between the same named entity. After that, different weights are given according to the importance of named entity, and a method of computing similarity between named entities is designed. Finally, according to the given threshold, to obtain whether planning and realistic matching, and the actual personnel working situation, provide the basis for personnel assessment for the management personnel.
1. INTRODUCTION
The performance of railway security supervisors includes: (1) performance projects and plans. According to the method of "monthly safety performance appraisal form for managers", managers first input post performance items into the system, and maintain monthly work plan and quantitative indicators according to performance items. (2) Realistic performance. When the manager inspects or finds problems on the spot, he chooses the corresponding work plan in the system to input realistic information. In such a long period of work, a large number of text data have been accumulated. It is of great significance to match the actual work of the analysts with the monthly plan. These data rely on manual analysis for a long time. Data analysis is usually not timely and unreliable because of the large amount of data and human factors. It has important research and practical significance to use text similarity computing technology based on text mining, automatically calculating the semantic similarity between two texts, and judge whether the plan and the reality match.

Text mining technology aims to express, understand and extract words, grammar, semantics and other information contained in unstructured text strings through computer technology. Through mining and analyzing the facts and implicit positions, viewpoints and values, it can infer the intention and purpose of the text generator[1]. Text similarity calculation is to calculate the similarity between two texts. There are three kinds of methods for calculating text similarity. One is the traditional calculation method. Character matching is used to determine whether the text is similar or not, but because of the general content of the monthly plan of railway security supervisors, the actual content is more detailed, and it is difficult to match the characters. Second, the character representation is mapped to vector space, using cosine similarity calculation or other distance calculation methods, such as TF-IDF, word 2vec. However, the existing algorithms remain in the representation similarity, and fail to achieve real semantic similarity. Third is based on the construction of semantic computation. Ontology-based semantic similarity calculation[2], Because the complexity of ontology construction, it is a relatively large project to build ontology or knowledge map based on the semantic similarity computation of [2].

Based on the analysis of the characteristics of data, this paper applies the method of named entity recognition and concept similarity calculation to realize text similarity calculation. Named Entity Recognition (NER) refers to the recognition of text entities with specific meanings, usually including names, places, institutions and so on. NER tasks include entity boundaries and entity types. It can be regarded as a classification problem, so we can use methods such as Bayesian model based on classification[3], Support Vector Machine (SVM)[4], Maximum Entropy (ME)[5]. It can also be considered as a serialization problem, so Hidden Markov Model (HMM)[6], Maximum Entropy Markov Model (MEMM) and Conditional Random Field (CRF) model can be used[7]. Lin B Y et al.[8] proposed that the word representation of multi-channel information and traditional CRF are embedded into LSTM network. The synthetic word contains character-level information and syntactic features. The model is similar to the model in this paper, but there are differences in word representation. The similarity calculation of the same named entity is based on the similarity calculation of words or phrases. The text similarity calculation based on words is based on the method of synonym word forest[9]. Chinese word similarity is based on semantic resource Hownet and concept similarity calculation method based on Hownet. These methods are based on a large corpus to calculate the similarity between words.

2. COMPUTATION OF TEXT SIMILARITY
In this paper, the process of calculating the similarity between plan and realistic content mainly includes text preprocessing, named entity recognition, concept similarity calculation and final comprehensive analysis. The overall process is shown in Figure 1.
Text preprocessing process. First of all, the original personnel performance plan and performance realism are divided into multiple tasks according to line change and period respectively. Each task is clearly numbered. According to the content characteristics of task record, the named entity extracted is defined as time, place, task and problems found. Then the named entity is labeled by BIOES.

Named entity recognition process. The process of named entity recognition is divided into two steps: model training and model application. Model training is based on the combination of BiLSTM and CRF. The tagged corpus is applied to the model and the model is trained to achieve high accuracy. The application of the model is to put the original performance plan and realistic content into the trained model separately, and automatically extract the time, place, person and key elements of the problem for the use of conceptual similarity.

Conceptual similarity calculation process. Conceptual similarity calculation is to calculate the similarity between the entity of performance plan and the corresponding entity of performance reality, and to get the entity similarity between the two texts. According to the importance of the entity, the corresponding weight of the entity is given. For example, in defined entities, time is less important, because it is written in the monthly plan and filled in the month's realism, so time is not so important. On the contrary, task and location are more important concepts, because the matching between tasks and locations is similar, it can be basically identified as a task, which can be considered as a match between the plan and reality. Finally, given the threshold, the similarity between texts is considered to be similar if the similarity is higher than 0.8. As shown in Figure 2

Comprehensive analysis process. After getting the similarity of the two texts, we can analyze the personnel performance data. The main analysis of the actual implementation of the staff's work plan.

3. NAMED ENTITY RECOGNITION MODEL

The neural network model in this paper consists of embedding layer, BiLSTM layer and CRF layer. Figure 3 shows the entire network structure. The embedding layer is applied to construct the feature representation of Chinese in each sentence. Then, the feature representation vectors are put into the BiLSTM layer. Therefore, a hidden state is obtained at BiLSTM level, which is considered as the character feature vector. In the last CRF layer, the final prediction label of the input sentence is decoded. The whole neural network is implemented by tensorflow, a deep learning framework.
3.1 Tagging Data
Generally speaking, supervised deep learning algorithms need a lot of labeled data. So before the experiment, we need to annotate a lot of text data. Firstly, four named entities are defined according to the analysis requirements and the text characteristics of personnel planning and realistic records. Entity categories are shown in Table 1. Secondly, in order to distinguish a complete entity, we use B (start), I (internal), O (other), E (End), S (single) tags and named entity definitions to tag data. B, I and E represent the beginning of a multi-character entity, the middle part of a multi-character entity, and the end of a multi-character entity, respectively. S means that there is only one character entity. O represents other characters, not entities we define. For example, the first character of a time entity is marked as B-TIM, the remaining characters are marked as I-TIM, and the last character is automatically marked as E-TIM by the program. The average length of sample sentences is about 30 characters.

| Entity name          | Annotation definition | English full name |
|----------------------|-----------------------|------------------|
| Time                 | TIM                   | Time             |
| Location             | LOC                   | Location         |
| Task                 | TAS                   | Task             |
| Discovered problems  | PRO                   | Problem          |

3.2 Character Embedding Layer
Character embedding layer is the high-dimensional representation of words. Characters are represented in two ways. One is to use genism based on Chinese wiki corpus to represent vectors. Each character is represented by a 100-dimensional vector. The second is the use of dictionary representation based on stuttering participle. If each character after word segmentation is composed of a single character, it is expressed as 0. The representation of two characters is 1 and 3. The representation of three characters is 1, 2 and 3. Define a 20-dimensional vector to represent each character vector. Then the representations of both modes are input into the BiLSTM layer.

3.3 Bidirectional Long-term and Short-term Memory Network
Long-term and short-term memory network (LSTM) is a special recurrent neural network (RNN). It can learn to rely on information for a long time, especially the relevance of information. Because it has three gates to protect and control neurons. They are forgetting gate, input gate and output gate. Forgotten Gate
controls what information can be discarded. Input gate control requires new information to be stored. The output gate controls the information that needs to be output. The specific workflow of the LSTM unit is shown in Figure 4.

\[
i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
h_t = o_t \cdot \tanh(C_t)
\]

\(i_t, f_t, o_t, C_t\) represent the unit states of the input gate, the forgetting gate and the output gate at time \(T\). \(x_t\) and \(h_t\) represent the vector of the input vector and the hidden layer at time \(t\). \(\sigma\) refers to sigmoid activation function. \(W_i\) and \(b_i\) represent the weight vector and the bias vector, respectively.

BiLSTM has two LSTMs. An LSTM processing sequence from left to right contains the past information of data, which is called forward hiding layer. Another LSTM processes data from right to left, containing future information of the data, called a backward hiding layer. So the final hidden state of BiLSTM is the connection between the front and back hidden layers, as follows.

\[
\bar{h}_i = \text{lstm}(x_i, \bar{h}_{i-1}) \\
\bar{h}_j = \text{lstm}(x_j, \bar{h}_{j-1}) \\
h_t = [\bar{h}_i; \bar{h}_j]
\]

3.4 Conditional Random Field Algorithms

In the final prediction sequence, it is necessary to consider the correlation between the final tags. For example, in the input sequence, I-TIM cannot be behind I-LOC. These restriction information can be learned automatically by CRF layer in the training process. So conditional random field model is added to the model. If we use the softmax layer to predict the final tag, these constraints may be broken.

4. CONCEPTUAL SIMILARITY CALCULATION

This paper uses the concept similarity calculation based on HowNet. This paper takes the concepts represented by Chinese and English words as descriptive objects, and reveals the relationship between concepts and concepts, as well as the attributes of concepts as the basic content of the knowledge base[11]. It is a concept consisting of "sememe". Its similarity calculation method is a calculation method based on world knowledge. "sememe" is the smallest unit to describe a concept. It describes each concept with a series of original meanings. HowNet uses more than 1500 sememes and is divided into 10 categories: Event, entity, attribute, aValue, quantity, qValue, SecondaryFeature, syntax, EventRole, EventFeatures. These sememes are divided into four groups. "Basic semantics" are used to describe the semantic characteristics of a single concept. “Grammatical sememe" is used to describe grammatical features. “Relational sememe” is used to describe the relationship between concepts and concepts. It also includes "Other sememe ".

Since the extracted words belong to notional words after extracting the named entity of the text, some of them are phrases, so there is a problem that the phrase or word is not included in the original
corpus of HowNet. Unlogged words need to be processed. The process of processing unlogged words is shown in Figure 5.

When the words or phrases to be calculated are not included in the HowNet corpus, the unlisted words are segmented into concepts to form a linear list of multiple concepts and organized in reverse order. For example, after the segmenting of "signal machine", it can be stored as: [machine]→[signal]. Take the last three concepts. Then the concepts in the concept list are merged in order. Between the two concepts, if the second concept is a function word, the first concept is taken directly, otherwise the two concepts are added together. Entities processed by unlisted words belong to notional words, so the similarity between notional words is only considered here.

The basic idea of calculating the similarity between notional words is that the overall similarity should be based on partial similarity. A complex whole is decomposed into parts, and the global similarity is obtained by calculating the similarity between parts. [11]

We divide the semantic expression of notional concept into four parts: independent sememe description, other independent sememe description, relational sememe description and symbolic sememe description.

Independent sememe description: Compute the similarity between two independent sememes, expressed as Sim(S1,S2)

\[ Sim(p_1, p_2) = \frac{2l}{d_1 + d_2} \quad (9) \]

P1 and P2 denote two sememes. \( l \) denotes the equal position of two sememes, and is the path length of A and A in the hierarchical system of sememes.

Mean operation is used to calculate the similarity among other independent sememes, relational sememes and symbolic sememes. Among them, S1 and S2 denote the two sememes computed.

\[ L = Max(len_{s1}, len_{s2}) \quad (10) \]

L denotes the maximum length of two sememes.

\[ Sim(S1, S2) = \frac{\sum_{i=0,j=0}^{L} Max(sim_i, sim_j)}{L} \quad (11) \]

Sim(S1,S2) denotes the similarity between two sememes. simi,simj denote the similarity between independent sememes in a set of sememes. Other independent meanings are denoted by Sim2(S1,S2). The relational sememe similarity is expressed by Sim3(S1,S2), and the symbolic sememe is expressed by Sim4(S1,S2).

Computing the similarity between two concepts is expressed as follows.
Among them $\beta_i(1 \leq i \leq 4)$, because of the different importance of Sim1 to Sim4 to the whole sentence, the first independent sememe $Sim_1$ reflects the most important feature of a concept, so the weight should be set to the maximum, generally above 0.5, and there are $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ and $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$, where we take $\beta_1 = 0.5$, $\beta_2 = 0.2$, $\beta_3 = 0.17$, $\beta_4 = 0.13$.

5. EXPERIMENTAL RESULTS

In the experiment, from July 2017 to January 2018, more than 60,000 data were used as the experimental data. Among them, 1000 were used as training samples and 200 were used as testing samples to train named entity models. Then the model is applied to all data and named entities are extracted automatically. The results are shown in Table 2. Conlleval CRF++ tool is used to calculate the accuracy and recall rate.

| Entity name         | Accuracy | Recall |
|---------------------|----------|--------|
| Time                | 91.78%   | 92.31% |
| Location            | 94.98%   | 95.46% |
| Task                | 93.10%   | 96.00% |
| Discovered problems | 72.62%   | 68.54% |

From the experimental results, we can find that "time", "place" and "task" have higher accuracy. This is because the content and format of these types of named entity records are more standardized. However, the accuracy of the "discovered problem" is not very high because of the diversity of its own problems and the different habits of personnel records.

The results of personnel performance plan and realistic matching calculated by named entity extraction and concept similarity are compared artificially, and the accuracy is 94%.

6. CONCLUSION

In this paper, we use classical application scenarios and techniques in text mining field to realize long text similarity calculation step by step, which has been well applied in railway personnel performance analysis. Named entity recognition itself is also a typical application scenario of text analysis. In the process of analysis, converting unstructured text data into structured data can provide more dimensions of mining personnel performance information.

REFERENCES

[1] Dou Zhicheng. Opportunities and Challenges of Text Big Data Analysis Technology [J]. Electronic Finance, 2015(11):59-61.
[2] Sun Haixia, Qian Qing, Cheng Ying. Review of Ontology-based Semantic Similarity Computing Methods [J]. Modern Library and Information Technology, 2010, 26(1):51-56.
[3] Yavuz S R, Küçük D, Yazıcı A. Named Entity Recognition in Turkish with Bayesian Learning and Hybrid Approaches[M]// Information Sciences and Systems 2013. Springer International Publishing, 2013:129-138.
[4] Ju Z, Wang J, Zhu F. Named Entity Recognition from Biomedical Text Using SVM[C]// Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on. IEEE, 2011:1-4.
[5] Konkol M, Konopik M. Maximum Entropy Named Entity Recognition for Czech Language[M]// Text, Speech and Dialogue. Springer Berlin Heidelberg, 2011:203-210.
[6] Zhao S. Named entity recognition in biomedical texts using an HMM model[C]// International Joint Workshop on Natural Language Processing in Biomedicine and ITS Applications. Association for Computational Linguistics, 2004:84-87.

[7] Zhang Y, Xu Z, Zhang T. Fusion of Multiple Features for Chinese Named Entity Recognition Based on CRF Model[J]. Journal of Computer Research & Development, 2008, 45(6):1004-1010.

[8] Lin B Y, Xu F, Luo Z, et al. Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media[C]. Proceedings of the 3rd Workshop on Noisy User-generated Text. 2017: 160-165

[9] Tian Jiule, Zhao Wei. Word similarity calculation method based on synonym word forest [J]. Journal of Jilin University (Information Science Edition), 2010, 28(6):602-608.

[10] Rui Xun, Xiaoning Ma, Lianbao Yang. Application of Natural Language Processing's key technology in Intelligent Railway [J]. Railway Computer Application, 2018, Vol.27(10):40-43.

[11] Liu Qun, Li Sujian. Word Semantic Similarity Computation Based on HowNet [J]. Chinese Computational Linguistics, 2002.