Extraction of Diplomatic International Cooperation Elements Based on Neural Networks and Domain Knowledge (Postprint)

Authors: Zhang Zijing, Wan Changxuan, Liu Dexi, Liu Yu, Liu Xiping, Jiang Tengjiao

Date: 2019-01-03T00:00:00+00:00

Abstract

To gain real-time insights into valuable information regarding international bilateral cooperation, intelligently and efficiently extracting international cooperation elements from Web-based diplomatic news is of paramount importance. This work abstracts the extraction of international cooperation elements as a problem akin to named entity recognition. First, it defines the connotation of international cooperation elements; second, it extracts rules embedded with domain knowledge; third, it proposes an international cooperation element extraction method for diplomatic news texts that integrates neural networks with domain knowledge; finally, it conducts comparative experiments on the same corpus with both neural network methods and its own rule combinations, and experimental results demonstrate that the proposed method achieves superior performance.

Full Text

Preamble

Vol. 37 No. 3
Application Research of Computers
ChinaXiv Cooperative Journal

Extraction of Diplomatic International Cooperation Elements Based on Neural Networks and Domain Knowledge

Zhang Zijinga,b, Wan Changxuana,b, Liu Dexia,b, Liu Yua,b, Liu Xipinga,b, Jiang Tengjiaoa,b

(a. School of Information Management; b. Key Laboratory of Data & Knowledge Engineering, Jiangxi University of Finance & Economics, Nanchang 330013,

China)

Abstract: To obtain valuable information from bilateral cooperation in real time, the intelligent and efficient extraction of international cooperation elements from Web diplomatic news is of utmost importance. This paper abstracts the extraction of international cooperation elements as a problem similar to named entity recognition. First, it defines the connotation of international cooperation elements. Second, it extracts rules that contain domain knowledge. Third, it proposes a method for extracting international cooperation elements from diplomatic news texts that combines neural networks with domain knowledge. Finally, comparing this method with neural network methods and rule-based combinations on the same corpus, experimental results demonstrate that the proposed method achieves superior performance.

Keywords: international cooperation elements; neural networks; sequence labeling; named entity recognition; Web diplomatic news

0 Introduction

The extraction of international cooperation elements is a subfield of natural language processing (NLP) research. The extraction method must intelligently identify international cooperation elements appearing in diplomatic news texts, such as "Belt and Road Initiative," deep processing of agricultural products, the Datka-Kemin power transmission project, the "Protocol on the Accession of China to the WTO," and the Stuttgart German-Chinese Friendship Association. Based on this, researchers can further explore China' s international cooperation industrial structure, common industries, emerging industries, advantageous industries, characteristic industries, industrial cooperation tendencies, cooperation effectiveness, industrial migration, and cooperation weaknesses, thereby enabling knowledge discovery of China' s international cooperation landscape. This provides real-time information services for Chinese enterprises going global and avoids blind expansion. With the development of online news, diplomatic news carried on the Web (referred to as Web diplomatic news) possesses characteristics of authenticity, authority, breadth, and timeliness. Extracting international cooperation elements through the window of China' s Web diplomatic news offers a new research channel and necessary technical support for knowledge discovery regarding China' s international cooperation.

This paper abstracts the problem of international cooperation element extraction as a task similar to named entity recognition (NER). The goal of traditional NER is to identify information units in unstructured text, including person names, location names, organization names, and numerical expressions (such as time, date, amount, and percentage). NER is regarded as a sequence labeling task in linguistics, with similar tasks including word segmentation, part-of-speech tagging, and machine translation.

Most traditional and well-performing sequence labeling models are linear statistical models, including Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [**?**, **?**]. Their effectiveness is influenced by manually constructed features and the characteristics of the dataset itself. For example, NER performance is affected by part-of-speech features in word segmentation results. However, due to the substantial workload and expertise required for manual feature construction, improvements to such methods have encountered bottlenecks. In recent years, to overcome the limitations of traditional models, nonlinear neural network (NN) models have been widely applied to NLP problems with the emergence of word vectors, achieving results comparable to traditional methods in NER tasks [**?**, **?**].

Through reading and analyzing China' s Web diplomatic news, this paper proposes the concept of international cooperation elements, analyzes their characteristics, and extracts rules containing domain knowledge. Based on this, it proposes a basic strategy for extracting international cooperation elements: First, obtain Chinese diplomatic news texts from the Web, perform word segmentation, and manually annotate the segmented sequences. Second, train a word vector model based on a dataset composed of Web diplomatic news and Chinese Wikipedia. Third, train an international cooperation element extraction model using the BiLSTM-CNNs-CRF neural network structure based on the word vector model to complete preliminary extraction. Finally, improve the extraction results by analyzing the preliminary results and leveraging domain knowledge such as external dictionaries and extracted rules.

1 Related Research

This section briefly introduces research progress related to neural networks and sequence labeling tasks in NLP studies. In NLP research, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks are widely applied. Studies in [**?**, **?**] show that CNN models can effectively extract morphological information (such as prefixes or suffixes) from letters in English words and encode it into neural network representations. [**?**] has demonstrated that although RNN models can theoretically capture long-distance dependencies, they encounter difficulties in practice due to gradient vanishing and gradient explosion problems.

To address the gradient vanishing problem, LSTM was proposed as a variant of RNN. Each LSTM unit uses three gates to protect and control the cell state: the input gate controls the magnitude of input, the forget gate controls the magnitude of previous memory state input, and the output gate controls the magnitude of final memory output. LSTM utilizes past dynamic temporal information and current input to predict the current moment. To make full use of temporal sequence information, [**?**] proposed Bi-LSTM (Bidirectional LSTM). Bi-LSTM has achieved good results in many tasks that require both past and

future context information. The basic idea of Bi-LSTM is to present each sequence as two separate hidden states forward and backward to capture past and future information respectively, then concatenate the two hidden states to form the final output.

Sequence labeling is a hot research topic in NLP with a wide range of applications, including NER, part-of-speech (POS) tagging, shallow parsing, and machine translation.

2.2 Characteristics of International Cooperation Elements

Although international cooperation element extraction shares similarities with NER, international cooperation elements themselves possess many characteristics that general named entities do not have. Specific analysis reveals:

a) International cooperation elements may contain punctuation or special symbols, which often serve as boundary markers for dividing these elements.

Example 6: To continue cooperation that has been ongoing for 35 years, both sides agreed to strengthen exchanges in the fields of legal metrology standards for quality, time, and temperature under the framework of the China-Germany Metrology Cooperation Agreement signed in April 2014 between the National Institute of Metrology of China and the German Federal Physical-Technical Institute.

In Example 6, "quality, time, temperature legal metrology standards" is collectively regarded as one instance of the industry category of international cooperation elements, containing two enumeration commas. In Example 7, industry, urbanization, and agriculture are regarded as three separate instances of the industry category, divided by enumeration commas.

b) The same international cooperation element may be classified into different categories in different contexts.

Example 8: At the sixth ministerial meeting of the China-Arab Cooperation Forum held in June this year, President Xi Jinping proposed the grand vision of jointly building the "Silk Road Economic Belt" and the "21st Century Maritime Silk Road" with Arab countries.

Example 9: Premier Li Keqiang stated that China' s initiative to build the 21st Century Maritime Silk Road aligns with Indonesia' s development strategy.

In Example 8, the "21st Century Maritime Silk Road" is regarded as an instance of the platform category. In Example 9, classifying the "21st Century Maritime Silk Road" as an instance of the planning category better fits the context.

c) International cooperation elements with larger connotations may contain elements with smaller connotations.

Example 10: Many African leaders have expressed their desire to expand China-Africa highway cooperation and build a highway network, which China actively supports and is willing to strengthen cooperation with African partners to gradually connect African highways into a network.

In Example 10, "highway" and "highway network" can be regarded as two instances of the industry category, where "highway" has a smaller connotation contained within the larger "highway network."

Due to these characteristics, the following situations may occur in extraction results:

a) Symmetrically appearing punctuation marks in extraction results do not appear in pairs, such as (), 《》, and double quotes. This phenomenon is referred to as the "non-standard extraction" problem.

b) The extraction result is only a subpart of the true result, and both share the same category classification. This phenomenon is called the "incomplete extraction" problem.

c) During preliminary extraction, an entire or partial international cooperation element may be split into two or more elements, with the split elements sharing the same category as the true result. This phenomenon is called "split extraction."

d) The extracted international cooperation element is completely correct but incorrectly classified, i.e., the "classification error" problem.

e) International cooperation elements in the true result do not appear in the extraction result, i.e., the "complete non-extraction" problem.

f) Conversely, international cooperation elements appearing in the extraction result are not actual elements, i.e., the "complete extraction error" problem.

---

3 Extraction of Domain Knowledge

To address the characteristics of international cooperation elements mentioned in Section 2.2, this paper identifies and extracts the following rules containing domain knowledge:

Rule 1: In extraction results, (), 《》, and double quotes must appear in pairs. If the extraction result is affected, corresponding adjustments must be made.

Rule 2: Start words and end words are words that appear at the boundaries of international cooperation elements but are not extracted in preliminary extraction. If extraction result r satisfying Rule 1 contains at least two words, supplement start word (B) on the left and end word (E) on the right. If BrE, Br, or rE appears in the original sentence, use the expanded result as the new

extraction result. If more than one type appears, select the longer one as the result.

To better describe Rule 4, we first define trigger words and boundary markers:

Definition 2: Trigger words are words or phrases that co-occur with or are contained within certain categories of international cooperation elements and can be used to determine the category of these elements. Based on their position, trigger words are divided into three types: those appearing on the left are called pre-trigger words, those on the right are post-trigger words, and those inside are internal trigger words.

Definition 3: Boundary markers are symbols or words that determine the co-occurrence window size between pre-trigger/post-trigger words and international cooperation elements, mostly punctuation marks, occasionally words in long sentences that can narrow the scope. The co-occurrence window refers to the span from the beginning of the boundary marker to the end of the international cooperation element (pre-window) or from the beginning of the element to the end of the boundary marker (post-window).

Rule 4: For each extraction result, search for pre-trigger words in its pre-window. If found, change the category of the extraction result; otherwise, retain the original result. The same applies to the post-window. Internal trigger words are searched within the international cooperation element. The priority (decision power) of the three types of trigger words, from high to low, is: internal trigger words, pre-trigger words, and post-trigger words.

Among these four rules, Rule 1 addresses the "non-standard extraction" problem in preliminary results, while Rules 2, 3, and 4 utilize domain knowledge to solve the "incomplete extraction," "split extraction," and "classification error" problems respectively.

Finally, for the "complete non-extraction" and "complete extraction error" problems, lacking necessary clues, they cannot be solved through domain knowledge based on preliminary results and thus fall outside the scope of this study.

---

4 Extraction Method for International Cooperation Elements

International cooperation element extraction shares similarities with sequence labeling tasks. However, compared with traditional sequence labeling tasks, international cooperation element extraction involves more categories with uneven length distributions, leaving room for improvement in results from general sequence labeling methods. In terms of categories, international cooperation elements have more categories classified by "semantics," making classification more difficult than traditional person/location name recognition. In terms of length, the distribution is uneven. For example, "quality, time, temperature

legal metrology standards" in Example 6 contains multiple words and punctuation marks. To address these issues, this paper' s extraction strategy is: first, employ a neural network model with excellent performance on sequence labeling tasks for preliminary extraction; then, use the preliminary results as input and optimize them using extracted domain knowledge rules to obtain final results.

4.1 Neural Network Layer Training

In the neural network layer, the CNNs-BiLSTM-CRF structure [**?**] is adopted to obtain preliminary sequence labeling results. The CNNs layer utilizes character vectors corresponding to each character in a word, combining them through CNNs to obtain a character representation of the word. The network structure is shown in [Figure 2: see original paper]. The BiLSTM layer takes as input the concatenation of each word' s word embedding and its character representation, outputting the labeling status for each word. For sequence labeling tasks, the CRF layer is helpful for considering correlations between neighboring labels and jointly decoding the optimal labeling chain for a given input sentence.

During training of the CNNs-BiLSTM-CRF model, the parameters used in this paper differ slightly from those in [**?**]. Specific parameter selections are shown in . Adam is an optimization algorithm using first-order derivatives for learning rate optimization. Such algorithm choices affect batch size selection at the order-of-magnitude level, and after comprehensive consideration, the batch size is set to 20. A gradient clipping value of -1 (less than 0) indicates no gradient clipping is used.

4.2 Optimization Based on Domain Knowledge

After analyzing preliminary results extracted by the neural network layer, this paper finds that the six types of problems mentioned in Section 2.2 still exist. Among them, the "non-standard extraction" problem has two causes: first, unpaired special symbols in the dataset itself; second, symbol loss during extraction. Such problems are underlying extraction errors that may affect subsequent optimization, so preliminary results must be normalized first. In normalization, special symbols in extraction results must appear synchronously with those in the dataset. That is, if special symbols appear in pairs in the dataset, they must also appear in pairs in extraction results; otherwise, no special symbols should appear. Normalization corresponds to Rule 1 in Section 3.

For the "incomplete extraction" and "split extraction" problems, although they have extracted components of the true results, such results are incomplete and not genuine international cooperation elements. However, precisely because they have extracted components of the true results, they provide opportunities to optimize preliminary results through "expansion" and "merging." The specific strategies for "expansion" and "merging" correspond to Rules 2 and 3 in Section 3.

After optimizing these two problems, new international cooperation elements

are generated, and "classification errors" may still exist among new and original elements. Therefore, this paper proposes corresponding optimization strategies corresponding to Rule 4 in Section 3.

5.2 Experimental Design and Results Analysis

This section first analyzes the impact scope of the domain knowledge extracted in Section 3 on the dataset, then conducts comparative experimental analysis of precision (P), recall (R), and F1-score for international cooperation element extraction between the proposed method and the CNNs-BiLSTM-CRF model [**?**].

1) Effectiveness of Domain Knowledge Extraction

The effectiveness of domain knowledge extraction is shown in , with analysis based on dataSet1. Rule 1 normalizes preliminary results, Rules 2 and 3 address extraction errors caused by "incomplete extraction" and "split extraction," and Rule 4 improves "classification errors."

Since all four rules are proposed to optimize neural network layer results, the premise is to ensure correct results already extracted by the neural network layer are not affected while accurately compensating for or improving errors as much as possible. Therefore, general rule evaluation standards are not applicable. This paper proposes the following two metrics to evaluate domain knowledge extraction effectiveness:

Discovery Rate = (Number of Real Problems / Number of Discovered Problems) × 100%
Correction Rate = (Number of Corrected Problems / Number of Discovered Problems) × 100%

The discovery rate measures whether rules can identify as many erroneous results as possible without affecting correct results. The correction rate evaluates whether discovered errors can be corrected as much as possible.

2) Extraction Effectiveness of International Cooperation Elements

The experimental process is as follows: First, train a 300-dimensional word vector using dataSet1 and dataSet2. Then, use the word vector, training set, and validation set as input to train an international cooperation element labeling model through the network model. Finally, evaluate the model' s performance on the test set, as shown in . The experimental results of the proposed method are shown in .

For comprehensive and intuitive comparative analysis, the experimental results of the proposed method and the CNNs-BiLSTM-CRF model (considering F1-score) are compared using a bar chart, as shown in [Figure 3: see original paper].

[Figure 3: see original paper] shows that the domain knowledge-based optimization strategy significantly improves the F1-score of the CNNs-BiLSTM-CRF model. The F1-score for the planning category increased by 8.49 percentage points (10.25% improvement). The industry category increased by 0.55 percentage points (0.62% improvement). The project category increased by 9.68 percentage points (14.30% improvement). The agreement category increased by 9.42 percentage points (12.37% improvement). The platform category increased by 5.45 percentage points (6.48% improvement).

For the industry category, domain knowledge has minimal impact, resulting in low optimization effectiveness for three main reasons: First, industry category elements have short word lengths, so the CNNs-BiLSTM-CRF model already performs well (F1-score of 89.40%). Second, the industry category has a large quantity in the dataset (65.97% of total), providing good training effectiveness. Third, the "expansion" and "merging" rules mainly optimize longer elements (requiring length 2 words).

For the project category, despite achieving the greatest F1-score improvement (14.30%) from domain knowledge, its post-optimization F1-score remains the lowest (77.37%) because the CNNs-BiLSTM-CRF model' s F1-score was already the lowest (67.69%). Two main reasons are identified: First, the model' s recall for project categories is notably low (R = 66.00%). Although domain knowledge discovers new elements and improves recall to some extent, the optimized recall remains low (R = 79.50%). Second, project category elements have more complex structures than other categories, resulting in low precision (P = 69.47%) that remains low after optimization (P = 75.36%).

For the platform category, most misclassifications involve organization names, which are also targets of traditional NER. As shown in , the platform category ranks second in proportion (16.01%). Therefore, the CNNs-BiLSTM-CRF model achieves excellent and balanced precision and recall (P = 83.02%, R = 85.13%). However, due to uneven length distribution, platform elements encounter more "incomplete extraction" and "split extraction" problems than traditional organization names.

For planning and agreement categories, these elements are generally longer and have small proportions in the dataset (3.32% and 3.90% respectively), resulting in poor extraction performance by the CNNs-BiLSTM-CRF model. Domain knowledge effectively improves F1-scores for these categories.

---

6 Conclusion

This paper first uses word segmentation tools to segment Web diplomatic news texts in the corpus, then employs neural networks for preliminary extraction of international cooperation elements, and finally combines manually extracted domain knowledge to optimize preliminary results and obtain final extraction

results. In the experimental phase, comparative analysis between the proposed method and neural network methods on the same corpus verifies that the proposed method achieves better results. The effectiveness of manually extracted domain knowledge on the corpus is also compared and analyzed.

Experimental analysis reveals that the proposed method has strong dependence on preliminary extraction results. Therefore, improving preliminary extraction performance is a future priority, especially for project category elements. Second, the distribution of the five categories of international cooperation elements in the corpus is highly skewed. Future work will consider expanding the dataset to construct one with more balanced category distribution for further experimentation and improvement. Finally, research will continue on knowledge discovery using extracted international cooperation elements from Web diplomatic news.

References

[1] Luo Gang, Huang Xiaojiang, Lin C Y, et al. Joint entity recognition and disambiguation [C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 879-888.

[2] Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution [C]// Proc of the 18th SIGNLL Conference on Computational Natural Language Learning. Stroudsburg, PA: ACL, 2014: 78-86.

[3] Hu Zhiting, Ma Xuezhe, Liu Zhengzhong, et al. Harnessing deep neural networks with logic rules [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers. Stroudsburg, PA: ACL, 2016: 2410-2420.

[4] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks [C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2013: 6645-6649.

[5] Zadrozny B. Santos C D, Zadrozny B. Learning character-level representations for part-of-speech tagging [C]// Proc of the 31st International Conference on Machine Learning. New York: ACM Press, 2014: 1818-1826.

[6] Chiu J, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association of Computational Linguistics, 2016, 4(1): 357-370.

[7] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C]//Proc of International Conference on Machine Learning. New York: ACM Press, 2013: 1310-1318.

[8] Dyer C, Ballesteros M, Ling Wang, et al. Transition-based dependency parsing with stack long short-term memory [C]//Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint

Conference on Natural Language Processing, Volume 1: Long Papers. Stroudsburg, PA: ACL, 2015: 334-343.

[9] Gael J V, Vlachos A, Ghahramani Z. The infinite HMM for unsupervised PoS tagging. [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2009: 522-530.

[10] Sha Fei, Pereira F. Shallow parsing with conditional random fields [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg, PA: ACL, 2003: 134-141.

[11] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C]//Proc of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 260-270.

[12] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [C]//Proc of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers. Stroudsburg, PA: ACL, 2016: 1064-1074.

[13] Chung J, Cho K, Bengio Y. A character-level decoder without explicit segmentation for neural machine translation [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016, Volume 1: Long Papers: 1693-1703.

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