An unsupervised learning method for named entity relation extraction of space knowledge graph

An unsupervised learning method for named entity relation extraction of space knowledge graph

Zhanji Wei[1*], Lingyong Huang[2], Gang Wan[1], Yao Mu[1] and Yunxia Yin[1]

[1] School of Space Information, Space Engineering University, Beijing, 101400, China

[2]China Centre for Resources Satellite Data and Application, Beijing 100094, China

[*]Corresponding author's e-mail: weizhanji@163.com

Abstract. Knowledge graph is a kind of typical artificial intelligence technology, which can improve cognitive ability of machines. Space is a booming domain for human beings. In this paper, space knowledge graph is proposed in order to improve intelligence of space information system. A major obstacle of space knowledge graph construction is to identify and extract core named entities and relations in space domain. An unsupervised learning method to extract core named entities and relations is proposed in this paper. Firstly, confidence level index used to measure importance of entities and conditional confidence level index used to measure credibility of relations between entities are defined and investigated. Then, an entity and relation extraction procedure based on maximum likelihood estimation and apriori algorithm is designed. Finally, a case study is carried out and influence of super parameter setting on the extraction result is discussed. The method proposed in this paper can be used to construct space knowledge graph at a lower cost compared to supervised method, and can be extended to other domain knowledge graph construction.

1. Introduction

Knowledge graph is a kind of coding technology for human knowledge. Since proposed by Google in 2012, knowledge graph has become more and more important in many fields. By expressing complex human knowledge in the form of triplets R(subject, predicate, object) interlinked one another, computers can better understand human knowledge, and on this basis, better analyze, calculate and reason, thus make the computer has more human-like intelligence. Knowledge graph has great application potential in intelligent retrieval, intelligent question answering, assistant decision-making and other fields, but the premise of these applications is to build the knowledge graph. At present, in the general field, the scale of Knowledge graph has reached 100 million level entities and relations, supporting applications such as Google and Baidu search engine. However, in the professional field, large-scale Knowledge graph is rare relatively.

At present, mankind has entered into space age, and the prospect of astronauts' long-term living and working on orbit is growing. In this paper, we propose to construct a space knowledge graph so that it can be used as a part of on-orbit spacecraft intelligent system to support applications such as virtual astronaut assistant.

Named entity and relation extraction is a procedure to get knowledge triplets from natural language texts, which is a key step to construct a space knowledge graph. A lot of researches on knowledge graph focus on the extraction of named entities and relations[1-3]. M. Miwa investigated relation extraction using recurrent neural network lstms[4]. X. Ma proposed to apply bi-directional lstm-cnns-crf model in named entity recognition and other sequence labeling tasks[5].

To extract entities and relations, supervised learning methods require a large number of training texts with correct answers to train a prediction model[6], and then the model can be used to get new instances of entity and relation. There are several problems in the application of supervised learning method in space domain. First, existing training texts are acquired from the general field, thus there are no available training texts in the field of space engineering. Even though the trained model performs well in the general field, the prediction effect deteriorates sharply if it is directly used in space domain, and it is almost unusable. Second, if the raw data in space domain is labeled to get training texts, a dilemma also exists that if the training set is too small, the model training is not sufficient, and the prediction effect is poor. However, extend the size of training set requires a lot of manpower and material resources. Worse still, even the model is trained with great costs, the portability will not be satisfactory since the effect of entity and relation extracting is poor for texts with different distribution patterns from the training set. Therefore, in this paper, we propose an unsupervised learning method to extract entity relations in space domain to circumvent drawback of supervised learning methods. The proposed method can mine and extract specific relational patterns from massive texts, and then use them to construct a space knowledge graph.

2. Methodology

An unsupervised learning method based upon law of probability statistics of natural language corpus in space domain is investigated thoroughly to mine features from the original texts, through which extraction of named entities and relations could be executed automatically. Apriori algorithm is studied and utilized to further accelerate the process of named entities and relations extraction from massive amounts of texts in space domain.

2.1. Formulation

Since the more and unbiased original data as is, the more accurate and credible will the probability statistics be, thus the first step to construct a space knowledge graph is to collect raw texts in space domain as many as possible. Sources include but not limited to official websites of space agencies, online scientific and technological literature databases and space theme websites. Tools such as scrapy are used in this process.

The second step is to cleanse the original texts to get a clean corpus for follow-up studies. In this step, sentences are seen as basic processing units. Cleansing procedure include text cutting, word segmentation and part of speech tagging, through which raw texts are transformed into a clean corpus set S_c and expressed in the form of key-value pairs, as is shown in the table 1. The last step is to traverse the corpus to explore statistical characteristics of different entities and relations in order to extract them.

Table 1. Form of the corpus S_c after cleansing of raw texts.

key	value
1	word1, word2,word3……
2	word1, word2,word3……
3	word1, word2,word3……

2.2. Confidence level index definition and calculation

Obviously, if there exists a certain relationship between two specific entities, the cooccurrence probability of words representing these entities and words representing their relationship in a sentence of natural language texts must be higher than that when they have no relationship. Therefore, the entity relationship contained in texts can be mined by calculating the joint probability of different words appearing together.

Define $S_{\ker nel}$ is a set of key words in space domain. The conditional probability of a word sequence $\mathbf{w_1^n} = \{w_1, w_2, \ldots\ldots, w_n\}$ in corpus S_c is $P\left(\mathbf{w_1^n} \mid S_c\right)$, and the prior probability of the word sequence is

$P(\mathbf{w}_1^n)$. The ratio of conditional probability and prior probability of $\mathbf{w}_1^n$ in the corpus is defined as the confidence level index $Con(\mathbf{w}_1^n)$:

$$Con(\mathbf{w}_1^n) = \frac{P(\mathbf{w}_1^n \mid S_c)}{P(\mathbf{w}_1^n)} \tag{1}$$

$P_{k-critic}$ is a pre-defined minimal confidence level index threshold. If $Con(\mathbf{w}_1^n)$ satisfies:

$$Con(\mathbf{w}_1^n) > P_{k-critic} \tag{2}$$

Then $\mathbf{w}_1^n \in S_{kernel}$.

Suppose that entity A and entity B has relation R. Entities A and B are composed of several words respectively:

$$A \Leftarrow \mathbf{w}_1^{n_1} \tag{3}$$

$$B \Leftarrow \mathbf{w}_1^{n_2} \tag{4}$$

Where symbol '$\Leftarrow$' denotes 'be represented by'. Relation R is also composed of several words. However, compared with entity A and B, the expression of relation R will be more abundant due to the diversity of natural language. Therefor, there may be many kinds of expressions of R:

$$R \Leftarrow \mathbf{w}_1^{r_1}, \mathbf{w}_1^{r_2}, \ldots, \mathbf{w}_1^{r_i}, \ldots, \mathbf{w}_1^{r_m} \tag{5}$$

The conditional confidence index of the ith expression of relation R between entities A and B is defined as follows:

$$Con(\mathbf{w}_1^{r_i} \mid \mathbf{w}_1^{n_1}, \mathbf{w}_1^{n_2}) = \frac{Con(\mathbf{w}_1^{n_1}, \mathbf{w}_1^{n_2}, \mathbf{w}_1^{r_i})}{Con(\mathbf{w}_1^{n_1}, \mathbf{w}_1^{n_2})} \tag{6}$$

$P_{c-critic}$ is a pre-defined minimal conditional confidence index threshold of relation R. If:

$$Con(\mathbf{w}_1^{r_i} \mid \mathbf{w}_1^{n_1}, \mathbf{w}_1^{n_2}) > P_{c-critic} \tag{7}$$

Then we can conclude that $\mathbf{w}_1^{r_i}$ represents a credible relation of entity A and B.

2.3. Apriori Algorithm

It can be seen from equations (1),(6) and (7) that one of the key steps to extract entities and relations in space domain is to acquire the conditional confidence level index of the relation R, which needs to calculate the joint probability of words appearing together with arbitrary number. According to the law of large number in probability theory, the maximum likelihood estimation method is able to calculate the joint probability of word sequence appearing together, which can be converted into counting their synchronized appearances in the corpus. Therefore, we need to traverse the corpus database for word frequency statistics.

Since the corpus is very large, if we do not apply any techniques, the amount of computation will increase exponentially when the number of words increases. For a sequence composed of n words, the number of computation will reach 2^n-1. Therefore, apriori algorithm is used to reduce the amount of calculation. Apriori algorithm adopts the idea of dynamic programming, which can greatly reduce the number of candidate word combinations that do not meet the threshold requirements and reduce the computational complexity. Apriori algorithm is based on two basic principles:

1) The subset of a frequent set must be frequent, i.e., if :

$$Con(\mathbf{w}_1^n) > P_{k-critic} \tag{8}$$

Then:

$$Con(\mathbf{w}_1^{n-1}) > P_{k-critic} \tag{9}$$

2) The superset of a non frequent set is necessarily not frequent, i.e., if :

$$Con(\mathbf{w}_1^n) < P_{k-critic} \tag{10}$$

Then:

$$Con\left(\mathbf{w}_1^{n+1}\right) < P_{k-critic} \tag{11}$$

According to above principles, many unnecessary calculations can be reduced. Take a word sequence set composed of four words {A, B, C, D} as an example. As is shown in the figure 1, if $Con(B) < P_{k-critic}$ has been proved, the confidence level index of all word sequence sets (shaded regions in the figure) containing word {B} need not be calculated.

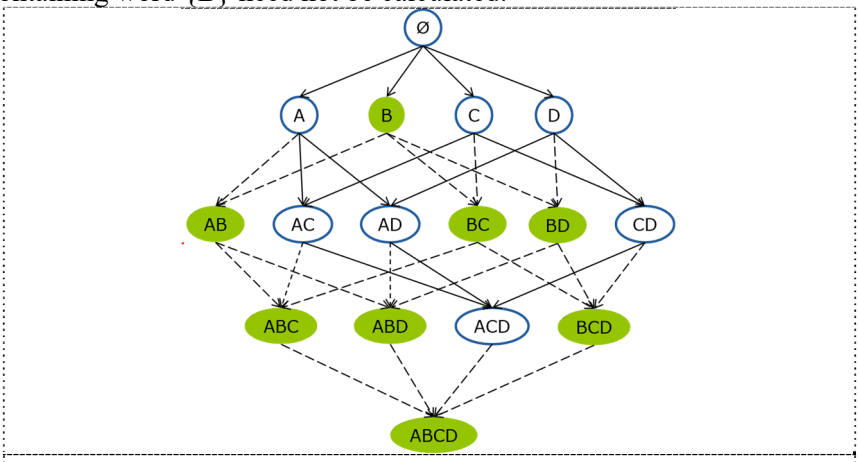

Figure 1. Schematic of the apriori algorithm.

The detailed calculation process applying apriori algorithm is as follows:

1) Let i = 1, scan the corpus S_c for the first time, and calculate the confidence level index of all the words. If the confidence level index of the words is greater than the threshold, it will be added to the set L (i), otherwise it will be discarded.

2) Let i = i + 1 and combine the words in the set L(i-1) to get the set C(i). Each element in the set C(i) is composed of i words. The confidence level index of each element in the set C(i) is calculated. If the confidence level index of an element is greater than the threshold, it is added to the set L(i), otherwise it is discarded.

3) Repeat step 2) until L(i) is an empty set.

4) According to equation(6)(7), credible entities relations can be found from L(i) and added to the knowledge graph.

3. Case study

Taking the knowledge triplet R(spacecraft, launcher, launch vehicle) representing relationship between spacecraft and launch vehicle in space domain as a study case, a large number of sentences shown in Table 2 can be retrieved and found in the original corpus.

Table 2. Original corpus in space domain.

ID	Sentences
1	Herschel was launched on an Ariane 5 rocket……
2	ISO was launched on an Ariane 44P rocket ……
3	COBE was launched on a Delta-5920 rocket……
4	The Hubble Space Telescope was launched on the Space Shuttle……

Because sentences show in Table 2 represent certain facts in space domain, they appear repeatedly in the original corpus. Applying the method expatiated in section 2, by calculating the frequency of occurrence and comparing with the predetermined threshold, the relations between different spacecraft entities and their launch vehicles will be found out without difficulty.

The two thresholds in equation(2)(7) are hyper parameters and must be determined from the start. As is shown in figure 2, with increase of the threshold, the precision rate of extraction will increases accordingly while the extraction rate will decreases, so some trade-off must be made between extraction rate and precision rate. There is an equilibrium point in figure 2, at which precision rate equals extraction rate. The equilibrium point can be found by experience. At the beginning, set the threshold at a relatively large value. If the extraction rate is too small, the threshold can be reduced repeatedly until the equilibrium point is reached.

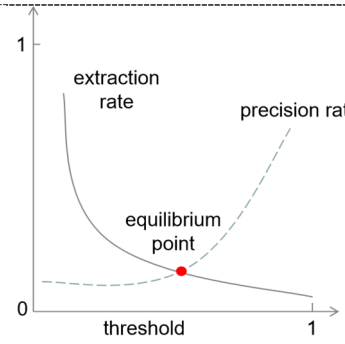

Figure 2. Trade-off between extraction rate and precision rate.

4. Conclusion

In this paper, an unsupervised learning method based on statistical law is proposed to extract named entity relations and construct space knowledge graph. Compared with supervised learning methods, the proposed method has an advantage that it does not need any labeled data to train the model. The method can find core vocabularies in space domain from raw texts by calculating the confidence level index of entities, and can extract relations between entities by calculating the conditional confidence level index of word sequences representing relations between entities. Apriori algorithm is also investigated to reduce unnecessary calculation, leading to improvement of the extraction efficiency. The proposed method can be extended to the construction of various other domain knowledge graph.

References

[1] Meishan Z., Yue Z., Guohong F. (2017) End-to-end neural relation extraction with global optimization. In: Proceedings of EMNLP. Copenhagen. pp. 1730-1740.

[2] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, Zhi Jin. (2015) Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of EMNLP. Lisbon. pp. 1785 – 1794.

[3] M. Peters, W. Ammar, C. Bhagavatula, R. Power. (2017) Semisupervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver. pp. 1756 – 1765.

[4] M. Miwa, M. Bansal. (2016) End-to-end relation extraction using lstms on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin. pp. 1105 – 1116.

[5] X. Ma, E. Hovy. (2016) End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin. pp. 1064 – 1074.

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