Research on Semantic Feature Disambiguation Based on Improved Collaborative Filtering Algorithm

Yu Linlin, Su Li, Lin Qiang
Dalian Institute of science and technology

Abstract. Artificial intelligence is developing rapidly, and the amount of information data is also increasing. Human-computer interaction technology can use computers to understand and process massive amounts of large data texts and accurately locate useful information. Moreover, it is very difficult to achieve smooth human-computer interaction. The main reason for the difficulty is that there are various ambiguities in natural language. Therefore, a noun phrase referential disambiguation model is established by improving the convolutional collaborative filtering algorithm in the paper, where the hand-crafted features are applied to clean up data to remove redundancy, and the remaining correct data is used as the input of the model. After experiments, the MACSSR algorithm and the CSSR algorithm are used for comparative testing under different M directs. It proves that the research content of this paper is superior to the comparative test method, and can extract semantic dependencies more quickly and effectively, which contributes to the intelligent semantic disambiguation technology.

1. Introduction

Natural language processing, the intersection subject of computer science and linguistics, aims to enable computers to understand human language, so that humans and computers can communicate with each other through natural language, and the gap between natural language and machine language can be bridged. What is more, in natural languages, words expressing different meanings in different language environments often appear, which are usually called ambiguous words or polysemous words [1-3].

Based on improved collaborative filtering algorithm and semantic ambiguity relation, a kind of convolution algorithm is proposed in the paper, which implements entity disambiguation for the collected scientific researcher entity data sets, extracts entity multi-primary attributes, successively calculates the structural semantic relationship between entities and the attribute classification semantic relationship, and finally obtains the similarity among entities. In addition, the convolution algorithm is used to realize entity disambiguation.

2. Analysis of Semantic Ambiguity Based on Feature Vector

2.1. Semantic Feature Screening of Convolutional Collaborative Filtering

Noun phrases refer to the disambiguation problem. Hand-crafted features is first used for preliminary screening to reduce the amount of data to obtain the best resolution effect in the paper, and then the teamed matrix is adopted as the input of the CNN. Finally, convolution is performed in same mode. Assuming the number of candidate antecedents after screening is $\alpha$, the number of anaphors after screening is $\beta$, and the word vector dimension is $\kappa$, the data scale will be $\alpha \cdot \beta \cdot \kappa$. Moreover, max-pooling is used to perform pooling, since the feature location information at this time is more important,
and average pooling will eliminate these location information. If the pooling layer size is $\beta \cdot \kappa$, a feature vector of $\alpha \cdot \kappa$ will be obtained [4,5].

CNN (Convolutional Neural Networks) that is adept at extracting spatial features is used to obtain deep semantic features. Meanwhile, the antecedent is vectorized as the input of BiLSTM, and BiLSTM that is good at time series problems is adopted to explore the implicit semantics between the context. After the combination of these features, a new combination feature will be formed, and the reference disambiguation task will be completed after softmax classification [6, 7]. The whole process is shown in Figure 1.

The extraction process of feature 1 is as follows. If the input data is matrix $\lambda$ whose dimension is $\phi$, and the number of words is $n$, then the width of the convolution kernel $\delta$ will also be $x$. Besides, if the height of the convolution kernel is $h$, the calculation expression of the feature matrix $F$ whose size is $n \cdot \kappa$ convolution layer after the same mode convolution operation is:

$$F = \text{Sigmoid}(\text{SUN} \cdot \lambda \cdot \delta) + \beta_1$$ (1)

The meaning of the expression is to use the convolution function SUN to convolve the matrices $\lambda$ and $\delta$, and add a bias term $\beta_1$. Finally, it is activated by the sigmoid function to obtain the feature dictionary $F$[8,9].

The extraction process of feature 2 is as follows. The model is a specially designed model. Compared with the traditional RNN model, the model proposed in the paper is better at remembering long-term information, since this model uses a uniquely designed "gate" to remove or add information to the cell state. Additionally, the model is composed of inputgate, outputgate and forgetgate three gates.

- Inputgate: $G_t = \zeta \left( \sigma^G + \sigma^G h_{t-1} + v^G \right)$ (2)
- Forgetgate: $J_t = \zeta \left( \sigma^J + \sigma^J h_{t-1} + v^J \right)$ (3)
- Outputgate: $U_t = \zeta \left( \sigma^U + \sigma^U h_{t-1} + v^U \right)$ (4)
- Memory unit: $\delta_t = G_t \cdot \text{anh}(\sigma_{w^a} + \sigma_{h^a} h_{t-1} + v_{\delta})$ (5)
- Output: $h_t = U_t \cdot \tan h_{\delta_t}$ (6)

In formulas (2)-(6), $\zeta$ refers to the activation function, $\sigma$ and $v$ are parameters. Inputgate enters the word vector at time $t$, and records which information needs to be updated. Forgetgate ignores some information, such as ignoring noun phrases that are not in the reference chain. Outputgate outputs the updated cell state [10].

**2.2. Hand-crafted Feature Extraction**

Features are data representations of the essential attributes of objects, and effective and accurate features can improve reference disambiguation [11].
(1) Anaphor PN: When the part of speech of the anaphor is a person's name, place name, organization name and other proper nouns, the feature value will be 1; otherwise, it will be 0. When tagging the corpus, part-of-speech tagging includes person names, place names, organization names and other proper nouns.

(2) Antecedent PN: When the part of speech of the antecedent is a proper noun, the feature value will be 1; otherwise, it will be 0.

(3) Anaphor Possession NP: If the anaphora is a noun phrase leading the genitive sign, the feature value will be 1; otherwise, it will be 0. During corpus labeling, the case category of nouns is marked, and feature extraction is judgment. If the genitive sign is led, the feature value will be 1; otherwise, it will be 0.

(4) Antecedent Possession NP: If the antecedent is a noun phrase leading the genitive sign, the feature value will be 1; otherwise, it will be 0.

(5) Anaphor Subject or object: If the anaphor is the subject or object, the feature value will be 1; otherwise, it will be 0. In addition, in the corpus tagging, the item is tagged with the syntactic structure. The subject is tagged as subject_subject, and the object is tagged as object_object.

(6) Antecedent Subject or object: If the antecedent is subject or object, the feature value will be 1; otherwise, it will be 0.

(7) Single and Plural Congruency: If the singular and plural numbers of the anaphor and antecedent are the same, the characteristic value will be 1; otherwise, it will be 0.

(8) Semantic Congruency: If the semantic categories of the anaphor and antecedent are the same, the feature value will be taken as 1; otherwise, it will be taken as 0. Additionally, the semantic categories are marked during the corpus labeling process. The semantic categories are shown in Table 2-1:

(9) Property Congruency: If the anaphor and antecedent have the same part of speech, the feature value will be 1; otherwise, it will be 0.

The semantic category is shown in Table 1.

| Numbering | Category | Numbering | Category |
|-----------|----------|-----------|----------|
| 1         | Hum      | 9         | rel      |
| 2         | substance | 10        | em       |
| 3         | food     | 11        | group    |
| 4         | event    | 12        | location |
| 5         | Loc      | 13        | time     |
| 6         | amount   | 14        | 0        |
| 7         | ani      | 15        | plant    |

3. Experiment Environment
Since the entity document needs to be cut first in the research, and then a matrix of the same scale will be established based on the cut out words to perform disambiguation calculations, it requires the computer running the algorithm to have a quantitative memory. Table 2 shows the specific experimental environment configuration.

| Configuration | Parameter |
|---------------|-----------|
| CPU           | (Intel) i7-9700K core 3.6GHz |
| RAM           | 128G DDR4 3200 |
| Hard disk     | (Intel) S4510 480G |
| operating system | Windows 10 Enterprise |
3.1. Experimental Data

In order to verify the effectiveness of the disambiguation algorithm which is improved and used in this study, it is necessary to compare the disambiguation algorithm with the referenced algorithm. 736 Doctors from Harvard University in the United States are selected as the search criteria in the paper to find a list of ambiguities of all names including Doctors on Google, and the list includes 6989 scholar entities. Meanwhile, the scholars in each group of ambiguity lists are sorted in descending order according to the number of citations, and then the first four are chosen as the experimental data set, which is a total of 1499. Additionally, if there are less than four scholars, they all will be retained.

Among the 1499 scholars, the number of Chinese papers collected through Google is 21186, where there are an average of 20 papers per person, and some scholars have less than 20 Chinese papers. Besides, the number of patents collected through the US Patent Office is 567, and there are 1052 personal basic information collected through Google in which some scholars’ personal web pages are missing.

Eventually, 736 Doctors scholars of the same name from Harvard University are used as the test data set to test the effectiveness of the algorithm. On the whole, there are a total of 736 groups of ambiguity lists, and each group has up to 4 entities with the same name. Therefore, the experimental data distribution is reasonable, which can test the accuracy of the improved disambiguation algorithm in the paper to a certain extent.

3.2. Algorithm Evaluation Standard

An improved collaborative filtering algorithm is adopted in the paper to establish a structured semantic relationship algorithm. What is more, MACSSR is an improvement based on the classification and structured semantic relationship algorithm CSSR. That is, on the basis of the original single-principal attribute classification, it is expanded to classification entity disambiguation with entity multi-principal attribute.

The commonly used algorithmic evaluation criteria precision (P), recall (R) and F value (F) are used in the paper to evaluate and compare the two. The specific calculation formulas are shown in (7), (8), (9).

\[ P = \frac{\theta}{\eta} \]  
\[ R = \frac{\theta}{\mu} \]  
\[ F = \frac{P \cdot R}{P + R} \]

In formulas (7)-(9), \( \mu \) is the total number of information in the sample, \( \eta \) refers to the total number of information extracted from the sample, and \( \theta \) represents the number of correct information.

In the paper, the accuracy rate represents the proportion of the number of correctly disambiguated entities in the test data set, and the recall rate refers to the proportion of the number of correctly disambiguated entities in the total number of entities in the data set. In addition, as a comprehensive index, the value of F is an evaluation index for integrating the two into consideration when there is a contradiction between P-value and R-value.

3.3. Experimental Process and Analysis

736 Doctors from Harvard University in the United States are adopted as the search condition in the paper to find 1499 name entities from the list of ambiguities including doctors on Google as the experimental data set and test data set. Besides, the single-attribute classification and structured semantic relationship algorithm CSSR and the MACSSR algorithm in the paper are respectively employed to
conduct comparative experiments under different $A$ values. The experimental results are shown in Table 3.

| $\gamma$ | CSSR P | CSSR R | CSSR F | MACSSR P | MACSSR R | MACSSR F |
|----------|----------|----------|----------|----------|----------|----------|
| 0        | 0.9134   | 0.9235   | 0.9346   | 0.9312   | 0.9642   | 0.9672   |
| 0.2      | 0.9342   | 0.9134   | 0.9023   | 0.9034   | 0.9434   | 0.9375   |
| 0.4      | 0.8437   | 0.8423   | 0.8431   | 0.82467  | 0.9034   | 0.9075   |
| 0.6      | 0.8131   | 0.8134   | 0.8164   | 0.8134   | 0.8985   | 0.8932   |

The following conclusions can be drawn from the table data.

When $\gamma$ is 0, the F value of the CSSR algorithm is 0.9346; when $\gamma$ is 0.2, compared with the traditional classification algorithm, the F value of the CSSR algorithm increases by 0.1%, which shows that CSSR performance has a certain improvement compared with traditional classification algorithms.

When $\gamma$ is 0.2, the F value of the MACSSR algorithm is the highest, which is 0.9672. What is more, compared with the traditional classification algorithm, the F value is increased by 1.9%, and the F value is increased by 1.8% compared with the CSSR algorithm, indicating that the performance of MACSSR is improved compared with the traditional classification algorithm and CSSR algorithm.

When the value range of $\gamma$ is (0.2, 0.6), as the number of people increases, the overall accuracy of the MACSSR algorithm shows a downward trend, which is in line with expectations.

After the first classification of entities, the similarity among the concepts of similar entities is relatively high. In order to further distinguish similar entities, more attention should be paid to the explicit semantic relationships between entities rather than implicit semantic relationships, that is, the value of $A$ should be reduced.

$\gamma$ is used to adjust the length of the node transformation in the structured semantic relationship. Generally, the smaller the $\gamma$ is, the smaller the influence of structuring on semantic relations will be, that is, the less implicit semantic relations are mined. Figure 6.1 shows the change of the F value of CSSR and MACSSR under different $\gamma$ values.
(1) Regardless of the value of the person, the F value of the MACSSR algorithm is greater than the F value of the CSSR, which shows that the overall performance of MACSSR algorithm is better than that of CSSR.

(2) On the whole, the larger the $\gamma$ value is, the smaller the F value of the CSSR and MACSSR algorithms will be, since the larger the $\gamma$ is, the more implicit semantic relationships among entities are discovered. Moreover, when the entities are classified for the first time through single-primary attributes or multi-primary attributes, the similarity among entities is relatively large. Meanwhile, too many implicit semantic relationships will hinder the further differentiation of entities.

4. Conclusion

(1) The algorithm in the paper is composed of two parts based on the improved collaborative filtering algorithm and semantic ambiguity relation algorithm, and $\gamma$ to some extent exists as the weight adjustment parameter of these two algorithms. The larger $\gamma$ is, the larger the proportion of the former in the overall algorithm will be, and the implicit semantic relationship between the mined entities will be more obvious as well.

(2) Using scholar entity data as the data source, when the character entity employs the character's gender, age and research field information as the multi-primary attributes for the first classification, there will be fewer similar entities and texts, and the description will be very similar. At this time, the larger the $\gamma$ is, the more difficult it will be to realize disambiguation. In addition, for different data sources, the classified entities and text ranges are not the same. Therefore, the utilization of $\gamma$ in the MASSR algorithm should be adaptively adjusted according to different data sources.

The problem of disambiguation is an important problem in the field of natural language processing. As a basic research task, it has a decisive influence on the research of natural language processing. Moreover, information from multiple data sources will be first manually disambiguated, then merged and stored in a local database. Finally, entity disambiguation will be studied, which will become the focus of future work, and further research will be carried out on real-time data collection.

Acknowledgement

This paper is supported by Research on Key Technologies of media data depth analysis system under the background of big data, which number is 2019-ZD-0349

References

[1] Mark-Christoph Müller, Florian Reitz, Nicolas Roy. Data sets for author name disambiguation: an empirical analysis and a new resource[J]. Scientometrics. 2017 (3)

[2] CHAI Mingke, LI Dongmei, ZHUANG Tingting, YANG Shuyi. Named Entity Disambiguation Based on Classified and Structural Semantic Relatedness[J]. Chinese Journal of Electronics. 2018(06)

[3] Wu Zhenxin, Shan Songyan. Comparison and Analysis of Typical Algorithms for Scientists' Correlation Measurement[J]. Digital Library Forum. 2019(03)

[4] Wang Wei, Li Xiaona, Yan Shuai. Research on author's co-authoring relationship prediction based on link analysis-taking the field of graphics and information as an example[J]. Modern Intelligence. 2018(11)

[5] Liu Junwan, Yang Bo, Wang Feifei. Research on Scholars' Influence Field Ranking Method Based on Citation Behavior and Academic Similarity[J]. Data Analysis and Knowledge Discovery. 2018(04)

[6] Zhang Bin, Li Yating, Dai Yiqing. Link Mining and Application Analysis of Disciplinary Cooperation Network[J]. Information Theory and Practice. 2018(09)

[7] Zhang Jinzhu, Yu Wenqian, Liu Jingjie, Wang Yue. Research on the prediction of scientific research cooperation based on network representation learning [J]. Journal of Information. 2018(02)
[8] Chen Erjing, Jiang Enbo. A review of text similarity calculation methods [J]. Data analysis and knowledge discovery. 2017(06)
[9] Zhu G, Iglesias C. Computing semantic similarity of concepts in knowledge graphs. IEEE Transactions on Knowledge and Data Engineering. 2017
[10] Zhang Libo, Sun Yihan, Luo Tiejian. A semantic similarity calculation method based on large-scale knowledge base[J]. Computer Research and Development. 2017(11)
[11] Liu Qiao, Zhong Yun, Liu Yao, Wu Zufeng, Qin Zhiguang. Integrated entity link algorithm based on semantic consistency [J]. Computer Research and Development. 2016(08)