Research on Entity Coreference Resolution Technology Oriented to Military Knowledge Graph

Ruijuan Hu*, Zhufeng Li and Jian Li
PLA Strategic Support Force Information Engineering University, Zhengzhou, China

*Corresponding author email: huruijuan01@126.com

Abstract. Data processing in military field is in the stage of fusion and disambiguation. The way of addressing the same thing is complex and difficult to integrate. Entity coreference resolution can effectively solve this problem. At the same time, entity coreference resolution is also an important part of knowledge fusion stage in the process of building military knowledge graph. This paper studies and implements Chinese coreference resolution from two aspects: Named Entity Recognition and coreference resolution. Firstly, use the BiLSTM+CRF model of neural network to realize NER in military field. By mining Wikipedia corpus, construct a pattern base, and iteratively find the coreference relationship in text based on pattern, and finally establish a model. To complete the rapid and effective construction of a total of 220,000 thesaurus, covering the military field of aircraft and ships two types of objectives, to achieve the military field entity coreference resolution, to provide strong support for the construction of military knowledge graph.

Keywords: Coreference resolution; Sequence annotation; NER; Pattern extraction.

1. Introduction
It is supposed to be the main problem in data processing that data source is extensive, data structure is complex, and data volume is huge. In order to optimize the internal structure of knowledge graph and integrate complete and more comprehensive entity information, effective entity link is an important issue to improve the quality of knowledge graph. At present, the data processing in the military field is in the stage of fusion and disambiguation. The way of addressing the same thing is complex and difficult. The entity coreference resolution can effectively solve this problem.

2. Coreference Resolution Technology
Coreference resolution is the process of merging different descriptions of the same entity in the real world. According to whether the coreferential relationship can leave the context, the coreference relationship can be divided into anaphora and coreference[1]. Anaphora refers to the close semantic relationship between the referent of the current text and the words, phrases or sentences that can be used as antecedents. The referential depends on the context semantic environment, and may not be established without the context referential relationship. The same referential may correspond to different antecedents in different language environments. Compared with anaphora, coreference mainly refers to two noun phrases pointing to the same entity in the real world, which can exist in a context out of semantic environment.

Chinese Coreference Resolution research started late, and because of the complexity of Chinese itself, the difficulty of coreference resolution is more difficult than that of foreign language. However, in recent decades, there have been different researchers in theory and technology implementation. The main research results are as follows: Shumin Shi proposed a method of domain named entity co deixis
digestion based on domain ontology semantic features and machine learning method[2]. Guochen Li and others used decision tree machine learning algorithm, combined with priority selection strategy[3]. Zhiqiang Wang proposed a two-candidate model of relevance antecedent when the current learning based reference resolution algorithms are all single candidate models[4]. Han Zhang proposed a hybrid method based on the BiLSTM-attention-CRF mode to solve the problem of coreference resolution in information security[5].

There are two main sources of data in the military knowledge graph: one is automatic collection from Wikipedia and other knowledge encyclopedia, the other is business data of military units. The data collected from Wikipedia shows the characteristics of many noise and large-scale data, which makes it difficult for users to obtain satisfactory query results quickly. Therefore, knowledge fusion after collection is required. Knowledge fusion includes two parts: entity link and knowledge merging. Entity link including entity disambiguation and coreference resolution, through which we can judge whether the entities with the same name in the knowledge base represent the same meaning, or whether the entities with different names can represent the same meaning[6]. This paper focuses on the characteristics of data noise and large scale in the military field, and the differences between Chinese and English coreference resolution.

3. The Process of Realizing Entity Coreference Resolution in Military Field

3.1. Entity Coreference Resolution Model

In order to achieve the entity coreference resolution in the military field, an approach base on patterns is used to iteratively find the coreference relationships in the text. Firstly, extract some coreference relationships from Wikipedia texts in advance, and then summarize the coreference patterns manually, establish the seed sets of patterns and coreference relationships, find the coreference pairs by iterating in Wikipedia, and expand the coreference pairs and pattern sets. The final results are presented in the form of coreference thesaurus. In the test and evaluation stage, the method effectively and accurately mine the coreference words of Wikipedia and military entities under the premise of ensuring the accuracy. The entity coreference resolution model is shown in Figure 1.

![Figure 1. The entity coreference resolution model based on patterns.](image)

3.2. Pre-mining Mode

3.2.1. Corpus preparation. Wikipedia is a multilingual encyclopedia collaboration project based on Wiki technology, which is written in multiple languages. Up to now, about 47 million items have been collected, enough to complete the task of coreference resolution. Download the offline version of Chinese Wikipedia, and preprocess it. It is necessary to convert the traditional to the simplified, extract the text of Wikipedia, stop words as well as segment Chinese words. Wikipedia XML documents and Infobox contain some coreference relationships and rules, which can be used to build initial seed sets. Using these seed sets, Wikipedia is used as the pattern extraction library to expand the pattern sets and
the coreference corpus. It is to get a relatively complete pattern sets, and prepare for later mining the coreference pairs in the military field.

3.2.2. Main Process of Pattern Mining. After the text of Wikipedia is divided into sentences, the words are segmented and the stop words are removed as the corpus, and then the skip gram model of word2vec is trained for standby.

- Using Python to analyse enormous Wikipedia's corpus. According to the characteristics of Wikipedia offline corpus, the coreference relationship between its title and Infobox is explored. According to these relations, a common thesaurus is established as the initial set (seed set) of the common thesaurus, and the patterns produced in the process of mining are recorded as the initial set (seed set).
- Extract the text of Wikipedia, for each text sentence, search the common reference pairs in the seed set. If a common reference pair is found, the pattern between the two words is saved to the candidate pattern set. Traverse pattern seed set.
- Use the N-gram model to extract the key words in the pattern set. N is 3 here.
- The Stanford-corenlp is used to identify and label the statement, the pattern is used to find the nearest two entities in the text, and the coreference relationship is put into the cluster. Finally, single cluster and empty cluster are deleted.
- After the first step of mining using the previous method, according to the candidate pattern set obtained in the previous step, n-gram model is used to calculate the weight of each word in the pattern, leaving keywords larger than the threshold as the effective pattern keywords.
- Start a new iteration, and use the common index seed set and pattern seed set obtained in the above steps to expand until the iteration threshold is reached or the convergence of the common index set is reached.
- Using the BiLSTM + CRF model to complete the sequence annotation of military corpus, we can extract the entities from the annotation effectively and accurately. Using the pattern set constructed above, we use the same iterative method to mine the coreference relationship in this field. Iterate mining until the collection no longer grows.

3.3. Named Entity Recognition in Military Field

Named entity recognition (NER) refers to the recognition of entities with specific meaning in text, which usually includes two parts: entity boundary recognition and entity category (person name, place name, organization name or other) [7]. The ultimate goal of this model is to achieve the coreference resolution of named entities. Currently, various named entity tools have a low recognition rate for this domain text. Therefore, sequence annotation is used to complete the recognition of named entities in this domain, and the recognized entities are taken as the basic objects of coreference resolution.

3.3.1. Determination of Data Annotation Type. The main technical methods of NER include: rule-based and dictionary-based methods, statistical based methods, hybrid methods, and neural network methods, etc. In recent years, deep learning has made breakthroughs in various fields. Neural network can effectively deal with many NLP task models [8]. We use a BiLSTM + CRF network built by tensorflow to annotate the text to identify the target entity in the text. Traditional NER tasks can generally define several types of entities: person name, place name and organization name [9]. According to the corpus, the following entities are identified:

| Type            | Annotation | Example                                      |
|-----------------|------------|----------------------------------------------|
| Person Name     | PER        | Mateush Moravitsky                           |
| Place Name      | LOC        | Santa Cruz Archipelago                       |
| Organization Name | ORG      | 998 Fleet formation                          |
| Aircraft        | AIR        | F-22                                         |
| Warship         | SHIP       | Wasp amphibious assault ship                 |
3.3.2. Corpus Annotations. The annotation adopts the format of BIO (entity start - entity internal - other), that is to say, for each entity, its boundary is determined by three labels of B, I and O, so as to find the entity in the text. In this paper, the word level annotation is used, and the BiLSTM model with memory function is used to complete the sequence annotation task. The data in the input layer will be calculated in two directions: forward and backward. The deep neural network LSTM performs the input feature representation, and softmax completes the sequence annotation based on the feature. For each entity, the first element is labelled as B-A, and other elements it contains are labelled as I-A. For entities that do not belong to any entity, such as, the input sequence = {Ticonderoga cruiser is a multi-purpose cruiser built by Ingels shipbuilding company}, and the annotated sequence = { B-JC I-JC I-JC I-JC I-JC O B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG O O O O O O O O O}.

3.4. Mining Criteria of Pattern Set
The reliability of the initial vocabulary will affect the mining of patterns and the construction of the final vocabulary. Firstly, the results of Wikipedia's XML and Infobox information analysis and extraction are put into the body of the entry to match, find out the phrases containing index words and common reference words, and extract the patterns through the N-gram model. In order to ensure the accuracy of pattern and coreference, a certain amount of manual summary should be added to the initial pattern set and coreference set. After getting the rule set, we use the rule to extend the co-reference pairs. According to the results of pattern location and NER, the best co-reference pairs are found around the pattern. In this way, through a certain number of iterations, all coreference pairs and pattern sets of Wikipedia are found.

The pattern set constructed by Wikipedia is a relatively complete pattern set, which limits the selection threshold of patterns and can get a relatively reliable pattern set. We use the pattern set from Wikipedia to mine the entity co-referential information in the corpus. Using the trained entity annotation model, we can get the best co-referential of each pattern, and then use the transitivity of the co-referential relationship to get a more complete co-referential relationship in the text. At the same time, we can get rid of the situation that only a single entity is mined.

The selection of the number of iterations needs to be combined with the corpus itself and the previous pattern set mining. The purpose of iterative mining is to comprehensively mine the co-referential relationship existing in the text, and try to get a more complete pattern. This paper chooses the number of iterations as 3, which has achieved good results.

4. Evaluation and Analysis

4.1. Evaluation Algorithm
The common evaluation algorithms of coreference resolution are MUC and B-CUBE. MUC evaluation algorithm is one of the earliest coreference resolution algorithms, which is a chain based evaluation standard algorithm and has been widely used in the evaluation task of MUC Conference. The disadvantage of MUC algorithm is that it does not consider the confidence of the common finger chain of a single entity, and treats all the wrong annotations equally. B-CUBE evaluation algorithm is an improvement of the MUC algorithm, which can overcome the shortcomings of the MUC algorithm. The B-CUBE algorithm calculates the recall rate and accuracy rate of each entity by calculating the correctly predicted entities in the entity chain. The final accuracy rate is calculated by accumulating weight * entity accuracy rate, and the final recall rate is calculated in the same way. This paper uses Precision, Recall and F1(F-measure).

\[
\text{Precision} = \frac{\text{Correctly identified coreference pairs}}{\text{Positive examples of model annotation}} \tag{1}
\]

\[
\text{Recall} = \frac{\text{Correctly identified coreference pairs}}{\text{Positive examples of corpus}} \tag{2}
\]

\[
F_{\text{measure}} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{3}
\]
F1 is a kind of measurement method that considers accuracy and recall comprehensively, also known as harmonic mean. For the task of entity recognition in this paper, we need to identify many kinds of entities, and use F1 macro average value to show its advantages and disadvantages. F1 macro mean value is the arithmetic mean of F1 values of various entities.

\[
F_{\text{ave}} = \frac{F_a + F_b}{2}
\]  

4.2. Coreference Resolution Process

4.2.1. Pre-mining mode. The mining of patterns plays an important role in the establishment of the final anaphor list. The above iterative mining algorithm is used to mine patterns. Determine \( n \) equals 3 for pattern selection, and select the first 20 pattern sets as the final pattern sets, and express 220000 coreference words in Wikipedia iteration.

4.2.2. NER. The model of BiLSTM + CRF is mainly used for sequence annotation in NER, 1400 of 2000 manually labelled data are used as training data, and 600 are used as training data. The recall rate, accuracy rate and F1 value of two kinds of important entities, aircraft and ship, are evaluated. It is a long and arduous process to adjust the parameters of the BiLSTM network. The data are divided into training data and test data, and different iterations and batch size are tried. The parameters of the network are adjusted according to the decrease of the loss function loss value and the annotation of the result F1 value. Due to the limited computer resources, it is impossible to use the smaller batch size for training. The decrease trend of loss in the final parameter scheme is shown in Figure 2. Under this parameter setting, the accuracy and recall rate of the two important targets are shown in Figure 3. Through the experimental verification, the F1 value of the two entities identified by the model has reached above 0.92, the F1 value of aircraft target recognition has reached 0.922, the F1 value of ship target recognition has reached 0.965, and the training effect is shown in Figure 4.

![Figure 2. Change of loss function.](image)

The F1 macro average value of two kinds of entities is 0.943, which is enough to support the next work of coreference resolution. From the results, although the accuracy of the two entities recognition has achieved the expected results, the recall rate of aircraft is still far behind the ship target. The reason is that the name of aircraft target is very simple, such as "F-22", while the name of ship target usually has a symbolic end, such as "Liaoning ship". In addition, BiLSTM has advantages in forming a long sequence. BiLSTM is more able to use sentence level semantic features for feature extraction. Therefore, the length of targets like ships is larger than that of aircraft, and the effect is better.
4.3. Evaluation and Analysis of Experimental Results

In the process of mining, the trained BiLSTM + CRF NER model is used for named entity recognition, and the mining pattern model is used for mining domain coreference pairs. In this paper, we mine two kinds of corpora, and evaluate the mining effect, that is, accuracy, recall and F1. The mining effect of the model is evaluated by calculating the number of coreference pairs and the number of actual recognition included in the corpus, and 500 corpora of two types are used for mining respectively. The results are as shown in Table 2.

| Entities  | Precision | Recall   | F1       |
|----------|-----------|----------|----------|
| Aircraft | 0.9694    | 0.9485   | 0.9325   |
| Warship  | 0.9280    | 0.89937  | 0.91347  |

Analysis of the experimental results shows that:
The main factor that restricts the mining of coreference pairs is named entity recognition. The recall rate of NER is not high enough, resulting in some empty clusters. The way to improve the model is to label new corpus constantly to improve the recall rate of NER. Although the entity recognition effect of the aircraft is slightly worse than that of the ship target, the mining effect is better than that of the aircraft target. The complexity of material description is related. The aircraft corpus itself has no hierarchical relationship between the model and the specific target, and the relative description is not
complicated with the ship target. Accuracy is one of the main advantages of the model. Both kinds of targets are mined on the premise of accuracy.

In this paper, we excavate coreference pairs from Wikipedia and military corpus, and finally obtain more than 220000 coreference pairs, which basically cover common entities in various fields, covering two kinds of targets in military field, namely, aircraft and warship. The proportion of word pairs mined in the military field is: 74% for the aircraft target, 25% for the ship target, and 1% for the organization.

References
[1] Yang Song et al 2020 *A Survey of Coreference Resolution Research Methods* (Beijing: Journal of Chinese Information Processing) pp 1-12
[2] Shumin Shi 2008 *Research on Chinese Coreference Resolution and related technologies based on Domain Ontology* (Nanjing: Nanjing University of Technology)
[3] Guochen Li et al 2005 *Chinese Pronominal Anaphora Resolution Via a Preference Selection Approach* (Beijing: Journal of Chinese Information Processing) pp 24-30
[4] ZhiQiang Wang 2006 *Research on the method of coreference resolution for knowledge graph* (Beijing: Beijing University of Posts and Telecommunications) pp 12-14
[5] Han Zhang et al 2020 *Research on coreference resolution technology of entity in information security* (Beijing: Journal on Communications) pp 165-175
[6] Wanying Sun 2018 *Research on the method of coreference resolution for knowledge graph* (Jilin: Jilin University) pp 35-36
[7] Jingbo Shang et al 2018 Automated phrased mining from massive text corpora vol 10, (IEEE Transactions on Knowledge and Data Engineering) pp 1825-1837
[8] Yuru Jiang et al 2020 *A Survey of Chinese Zero Anaphora Resolution* (Beijing: Journal of Chinese Information Processing) pp 1-12
[9] Fang Kong 2012 *Chinese and English pronoun resolution based on tree kernel function* (Hohhot: Journal of software) pp 108-109