Anaphora Ambiguity Detection Method Based on Cross-domain Pronoun Substitution

Fengyong Peng[†], Xi Wu[‡], Yongxin Zhao*[†§], and Yongjian Li[§]
[†] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, China
[‡] The University of Sydney, Australia
[§] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China

Abstract—**Pronoun anaphora ambiguity is very common in natural language descriptions, especially in specilized fileds such as computing, medicine and aerospace. When multiple antecedents appear before a pronoun word, readers with different background knowledge often have completely different understandings on a same word. In order to reduce such misunderstandings caused by ambiguity in the process of document propogation, we usually use manual methods to check the ambiguity of reference, which however cannot meet the increasing needs of detection with the development of various disciplines. In this paper, we propose a method to intelligently detect sentences with anaphora ambiguity. First of all, we identify criteria for ambiguous sentences and then use word embeddings to further detect ambiguity. Specifically, we propose a pronoun substitution strategy based on coreference resolution, and combine this strategy with word embedding techniques to generate a cross-domain anaphora ambiguity detection method. Finally, we carry out experiments on aerospace documents, which verify the effectiveness of our proposed method in anaphora ambiguity detection.**

Index Terms—**ambiguity detection, anaphora ambiguity, pronoun substitution, Cross-domain, natural language processing**

I. Introduction

There are a large number of documents closely related to professional background knowledge in various subject fields. In the process of document transmission, there are usually differences in understanding due to the different knowledge background of readers. We call such differences in understanding cross-domain ambiguity. The cost of manual ambiguity detection is getting higher and higher, and it has become an urgent problem to use intelligent technology to detect ambiguity in documents in specific fields.

In the field of natural language processing, the purpose of ambiguity detection is to first set an ambiguity standard and then judge whether a sentence can be interpreted in multiple ways under this standard. Ferrari $et\ al.$ have successively proposed the domain-specific ambiguity detection method based on word embedding [1], the cross-domain ambiguity identification method based on language model [2], and the cross-domain ambiguity detection method oriented to requirements engineering [3]. On this basis, Vaibhav Jain $et\ al.$ proposed cross-domain ambiguity detection using linear transformation of word embedding spaces [4], the word vector model is trained on the corpus of different fields, and then the word vectors obtained from different fields are mapped to the same space by using linear transformations. The ambiguity of words is measured by the difference of word vectors. Siba Mishra $et\ al.$ [5] have shown, through extensive and detailed experiments on several different subdomains, that word embedding based techniques are very effective in identifying domain-specific ambiguous words. The above studies have achieved good results in cross-domain ambiguity detection, but the above schemes also have some shortcomings. They only solve the problem of cross-domain ambiguous words without exploring the ambiguity of sentences in depth.

Inspired by the above work, we propose anaphora ambiguity detection method based on cross-domain pronoun substitution. The proposed method is tested in the aerospace field. Experimental results verify the effectiveness of the proposed method in anaphora ambiguity detection. Our major contributions include the following two points:

1) Our method combines pronoun substitution with cross-domain ambiguity detection and achieves good results in anphora ambiguity detection.
2) The cross-domain ambiguity detection method based on word embedding achieves good results at the word level. We consider the weight of words, design the construction method of sentence embedding, and realize the ambiguity detection at the sentence level.

The rest of this paper is structured as follows: Section II discusses related work; Section III describes the research methods; Section IV shows the result through experiment; Section V concludes our work and looks forward to future works.

II. Related work

The ambiguity of pronoun anaphora is a typical kind of ambiguity in natural language processing. We assume that there are multiple alternative antecedents before the pronoun, so we can think that the sentence will produce anaphora ambiguity when there is little difference between the probability of multiple alternative antecedents and the pronoun association. Based on the above assumptions, we use the clustering results obtained by coreference resolution to construct new sentences using the strategy of pronoun substitution, and detect sentence ambiguity by quantifying the difference of sentences after cross-domain pronoun substitution. The construction of the ambiguity detection method involves the following:

*Corresponding author: yxzhao@sei.ecnu.edu.cn

DOI reference number: 10.18293/SEKE2023-173

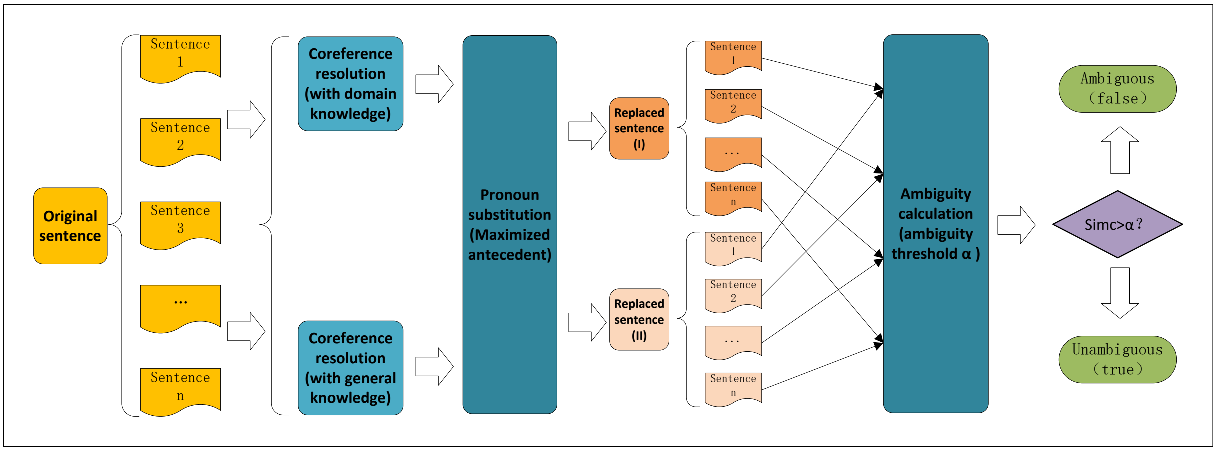

Fig. 1. Method architecture

A. Coreference resolution

In 2017, Kenton Lee *et al.* proposed end-to-end neural coreference resolution [6]. Subsequently, Kenton Lee *et al.* introduced a fully differentiable approximation of higher-order reasoning for coreference resolution [7]. In 2019, Mandar Joshi *et al.* proposed to apply $BERT$ to coreference resolution [8].Then Mandar Joshi *et al.* proposed that improving pre-training by representing and predicting spans [9] has achieved better coreference resolution results. The technology of coreference resolution is quite mature at present, and our research is carried out on this basis.

B. Embedding spaces alignment

Linear transformations can be used to learn linear mappings from one vector space to another. Mikolov *et al.* [10] used a set of word pairs $\{x_i, y_i\}_{i=1}^{n}$, in which y_i is the translation of x_i. Then they learn the translation matrix W by minimizing the following loss function:

$$\sum_{i=1}^{n} |x_i W - y_i| \tag{1}$$

This method can also be used to align monolingual word embeds. If the meaning of most words is assumed to remain the same, linear regression can be used to find the best rotation alignment between two words embedding spaces. This can be used to identify changes in meaning if words are not properly aligned. This is the basis of the proposed method for identifying cross-domain ambiguous words.

C. Text similarity

Sentence similarity is a complex phenomenon. The meaning of a sentence depends not only on the words in the sentence, but also on the way they are combined. There are also many dimensions of semantic similarity, and sentences may be similar in one dimension and different in others. Smooth Inverse Frequency (SIF) method is a reasonable method for sentence similarity calculation. The SIF method introduces a weight mechanism to reduce the weight of some unimportant words, while retaining the information that contributes more to the semantics. Our paper researches on this basis.

III. Anaphora ambiguity detection

The anaphora ambiguity detection method is shown in Figure 1, which mainly includes the following three parts:

A. Construct the coreference resolution model with different knowledge background

To judge whether a sentence is ambiguous, the first problem to solve is to check whether there are ambiguous concepts or polysemous words in the sentence. The space field is very broad, and the background knowledge of the participants is also very different. If the description of fuzzy concept or polysemous word appears in aerospace software carrier, people with different background knowledge are likely to have different understanding. We call such ambiguity phenomenon cross-domain ambiguity, that is, the ambiguity phenomenon caused by the different background knowledge of readers.

Vaibhav Jain *et al.* [4] showed that word ambiguities can be found by comparing word embeddings across domains. After being applied to the field of coreference resolution, Bert has achieved very good results in coreference resolution. Based on this, we propose to use different background knowledge to construct coreference resolution models. The improving pre-training by representing and predicting spans model proposed by Mandar Joshi *et al.* [9] has achieved the best results in practical applications. On this basis, we use the corpus from different fields to perform fine-turn operation on the bert pre-trained model to obtain coreference resolution models with different knowledge backgrounds.

B. A pronoun substitution strategy based on maximizing antecedents

Ambiguities caused by unclear references are common in the space industry. Coreference resolution can effectively solve the problem of unclear anaphora in text, but as a clustering operation, coreference resolution cannot be directly used for ambiguity detection. Here, we propose the strategy of pronoun substitution based on coreference resolution, that is, all the pronouns in the sentence are replaced with antecedents to get a new sentence without pronouns.

We assume that there exists a Chinese sentence S, $S = \{w_1\ w_2\ w_3\ w_4, w_5\ w_6\ w_7\ w_8.\}$. Where w_5 is a pronoun, w_2 and w_4 are nouns, and w_5 refers to w_2. We assume that w_2 is the phrase that can be segmented. So w_5 could point to w_2,w_{2-1},w_{2-2}, or w_4.If we apply the maximization antecedent algorithm 1, then w_5 cannot point to w_{2-1} and w_{2-2}.

Algorithm 1 pronoun replacement algorithm based on maximized antecedents

```
Input: Original sentence sen,Sentence segmentation results(full mode) $words = \{w_1, w_2, \ldots, w_n\}$,Coreference resolution result $clusters = \{c_1, c_2, \ldots, c_m\}$ and $pos = \{p_1, p_2, \ldots, p_m\}$.
Output: The new sentence after the substitution.
set $max\_w = ""$
for $i = 1 \to n$ do
    if $(c_1\ in\ w_i\ \&\&\ w_i\ is\ noun\ \&\&\ length(w_i) > length(max\_w))$ then
        $max\_w = w_i$
    end if
end for
$newSen = sen$
for $i = m \to 2$ do
    $replace\ newSen[p_i, p_i + length(c_i)]\ with\ max\_w$
end for
return $newSen$
```

C. Cross-domain ambiguity calculation method based on SIF

In this paper, ambiguity detection is carried out from the sentence level. Firstly, whether there is an anaphora in the sentence is judged. If there is a pronoun case, the pronoun substitution is carried out, otherwise no processing is done. During pronoun substitution, we obtained substitution sentences produced by models with different knowledge backgrounds. Then word Embedding is done on sentences with different knowledge backgrounds. Then, linear variation is used to bring the word embedding models obtained from different domains into a unified word embedding space. In a unified space, we detect ambiguity by analyzing the semantic similarity of two sentences. We believe that there is no ambiguity in the case of basically the same semantics, and there is ambiguity in the case of large semantic differences. Considering that each word in the sentence has different weight for the meaning of the sentence, we use the SIF method to calculate the sentence similarity. Although the traditional SIF method considers the influence of word weight on sentence meaning, it will bring new problems after the introduction of pronoun substitution. We assume that n pronouns have been replaced in the sentence, and the size of n will change the weight of the antecedent in the sentence. In order to eliminate the weight change caused by pronoun substitution errors, we adjust the sentence vector of SIF algorithm and obtain the improved semantic similarity calculation method:

$$\text{vector}(sen) = \sum_{w \in sen} \text{weight}(w) \cdot IDF_D(w) \cdot \text{vector}(w) \quad (2)$$

$$\text{weight}(w) = \begin{cases} \frac{1}{|sen|}, w \text{ for original word.} \\ \frac{1}{num \cdot |sen|}, w \text{ for substituted word.} \end{cases} \quad (3)$$

We assume that there is a threshold value α, when $Simc(sen_1, sen_2) > \alpha$, the sentence is considered ambiguous. When $Simc(sen_1, sen_2) <= \alpha$, sentence semantic similarity is considered to have no ambiguity.

IV. Experiment and analysis

A. Data

We use web crawler to obtain Chinese space corpus data from Wikipedia and clean the data. In total, we obtained $20,000$ articles, $9,132,781$ sentences, and $159,288$ words.

B. Parameter setting

Coreference resolution We extend the original tensorflow implementation of c2f-coref and bert using the approach of Joshi *et al.*[9] According to the experimental results of joshi *et al.*, Bert-large model was adopted, and set the $max\_segment\_len = 384$.

Word Embedding We use the gensim implementation of the word2vec SGNS algorithm for word embedding training. These hyperparameters are the same as those used by Jain *et al.* [4] in ambiguity detection.

C. Experimental results

Ambiguity calculation results We select 20000 statements from the aerospace software carrier to test the algorithm. In the first step, the corpus is input into the coreference resolution model with different background knowledge, and the anaphora chain and the corresponding position information are output if the sentence has an anaphora. In the second step, based on the original sentence and the anaphora chain (including location information), the pronoun substitution operation is carried out by maximizing the antecedent strategy. In the third step, the new sentences obtained in the second step are vectorized under word embedding models with different backgrounds, and then the word embeddings from different Spaces are mapped into the same space using linear spatial variation. In the fourth step, in the same word embedding space, the word vectors are accumulated according to the designed weights (formula) to obtain the sentence vectors, and then the cosine similarity of the two sentence vectors is calculated. Finally, the calculation results of ambiguity are shown in the Table I.

Human evaluation results We randomly selected 2000 statements from the 20000 statements that participated in ambiguity computation for manual evaluation. We select two groups of personnel with different background knowledge, one group is the technical personnel in the field of aerospace, and the other group is the administrative staff with no technical background in the field of aerospace. On this basis, the two

TABLE I
Ambiguity detection results

No.	Similarity	$\alpha = 0.4$	$\alpha = 0.8$	Manual
1	0.3996	ambiguous	ambiguous	ambiguous
2	0.9554	unambiguous	unambiguous	unambiguous
...	...	...	...	...
446	0.7364	unambiguous	ambiguous	ambiguous
447	0.6763	unambiguous	ambiguous	unambiguous

groups respectively performed pronoun substitution on 2000 statements according to the pronoun substitution strategy of maximizing the antecedent. Then, the sentence is marked by comparing the results of the two substitutions. If the results of the two substitutions are consistent, it can be considered that the sentence does not have the ambiguity caused by the ambiguity of anaphora. The manual evaluation found anaphora in 447 statements, of which 42 statements had cross-domain ambiguity. We correlate the human standard results and the results of the automated algorithm for detection as shown in the table I.

D. Discussion of ambiguity threshold

We assume that there is an ambiguity threshold α, and when the result of the sentence ambiguity calculation is greater than α, we can consider the sentence to be ambiguity free. When the result of the sentence ambiguity calculation is less than or equal to α, we consider the sentence to be ambiguous. We propose two measures in ambiguity threshold detection. Accuracy of ambiguity detection and coverage of ambiguous sentences. We use the automated ambiguity detection results and the human evaluation results to plot the ambiguity threshold versus the ambiguity detection accuracy and the ambiguity sentence coverage.

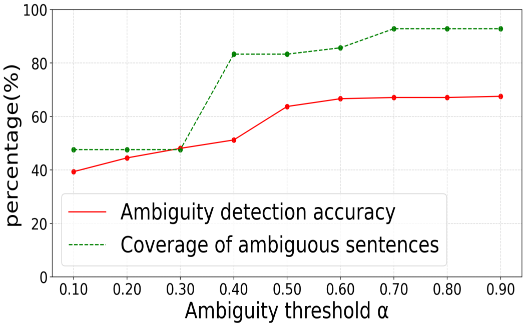

Fig. 2. Coverage and accuracy

By analyzing the images, we think that it is better to set the ambiguity threshold at least 0.7, at this time, it can almost cover most of the ambiguous sentences, and the accuracy of ambiguity detection can also be guaranteed. When the threshold α is 0.7, the ambiguity detection accuracy of the algorithm is 67.11%, and the ambiguity sentence coverage rate is 92.86%.

V. Conclution and future wiok

This paper proposes an intelligent method for anaphora ambiguity detection. Based on the research of cross-domain ambiguity and coreference resolution, we propose a cross-domain anaphora ambiguity detection method based on cross-domain pronoun substitution. Experiments in the space field verify the effectiveness of the method.

Domain-specific ambiguity detection is a very complex project. Subsequent work will continue to summarize the rules of other types of ambiguity sentences and study ambiguity detection methods with a wider scope of application.

VI. Acknowledgement

This work is supported by National Key Research and Development Program (2020AAA0107800), National Natural Science Foundation of China (NSFC 62272165), the “Digital Silk Road” Shanghai International Joint Lab of Trustworthy Intelligent Software (Grant No. 22510750100), and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

[1] Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. “Detecting Domain-Specific Ambiguities: An NLP Approach Based on Wikipedia Crawling and Word Embeddings”. In: IEEE Computer Society, 2017.

[2] Alessio Ferrari, Andrea Esuli, and Stefania Gnesi. “Identification of Cross-Domain Ambiguity with Language Models”. In: IEEE, 2018.

[3] Alessio Ferrari and Andrea Esuli. “An NLP approach for cross-domain ambiguity detection in requirements engineering”. In: *Autom. Softw. Eng.* 26 (2019).

[4] Vaibhav Jain et al. “Cross-Domain Ambiguity Detection using Linear Transformation of Word Embedding Spaces”. In: vol. 2584. CEUR-WS.org, 2020.

[5] Siba Mishra and Arpit Sharma. “On the Use of Word Embeddings for Identifying Domain Specific Ambiguities in Requirements”. In: IEEE, 2019.

[6] Kenton Lee et al. “End-to-end Neural Coreference Resolution”. In: Association for Computational Linguistics, 2017.

[7] Kenton Lee, Luheng He, and Luke Zettlemoyer. “Higher-Order Coreference Resolution with Coarse-to-Fine Inference”. In: Association for Computational Linguistics, 2018.

[8] Mandar Joshi et al. “BERT for Coreference Resolution: Baselines and Analysis”. In: Association for Computational Linguistics, 2019.

[9] Mandar Joshi et al. “SpanBERT: Improving Pre-training by Representing and Predicting Spans”. In: *Trans. Assoc. Comput. Linguistics* 8 (2020).

[10] Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. “Exploiting Similarities among Languages for Machine Translation”. In: abs/1309.4168 (2013).