Indian Journal of Science and Technology, Vol 8(21), DOI: 10.17485/ijst/2015/v8i21/78270, September 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Resolution of Ellipses in Coordinate Noun Phrases

Young Hee Jung[1], Kinam Park[2], Jeong Min Chae[1*] and Soon Young Jung[1]

[1]Department of Computer Science of Education, Korea University, Seoul, Korea; onacloud@korea.ac.kr
[2]Department of Computer Software Engineering, Soonchunhyang University, Korea

Abstract

A coordinate noun phrases connects two words or phrases together via a coordinate conjunction. The duplicate words of conjunctions are mainly omitted. This structure often occurs in the science literature. And this makes it difficult to understand of the sentence. Our research is motivated by the need to reduce the costs of misunderstandings that can occur during NER. We propose a method for resolving coordinate noun phrases with simple or complex ellipses using rules and dataset. And we describe a method to automatically build dataset. This method is applicable to a general-purpose in various fields. Our dataset effectively is used to distinguish between high and low modifier attachment. We reported on a set of experimental results to evaluate the performance of our approach. The results show that our system can efficiently resolve coordinate noun phrases. And we are sure that the method can resolve ellipses in various domains.

Keywords: Coordinate Noun Phrases, Named Entity Recognition, NP Ellipsis Resolution, Text Mining

1. Introduction

Numerous articles related to the NER are constantly being published. Accuracy performance of the Named Entity Recognition (NER) is also more enhanced. But there are some problems that are unresolved. One such problem is the ambiguity that occurs in coordinate noun phrases. A coordinate noun phrases connects two words or phrases together via a coordinate conjunction (e.g., 'and', 'or', etc). Consider the following real examples.

E1: Cytokine and immune receptor.
E2: IL - 1 beta, IL - 6 or TNF alpha mRNA level.

The duplicate words of conjunctions are mainly omitted. This structure often occurs in the science literature. And this makes it difficult to understand of the sentence. It is also not easy to accurately analyze by the syntax analysis. In particular, this structure makes it a little too difficult for NER. Something can be interpreted in more than one way.

In the example E1, the coordination construction can be bracketed as '[[cytokine and immune] receptor]' that is interpreted as 'cytokine receptor' and 'immune receptor' , or as '[cytokine] and [immune receptor]', in which case it is only 'immune' which is modified by 'receptor'. It is not easy to distinguish between the omitted words and original words. In the example E2, it is a complex noun phrases and includes three noun phrases: 'IL - 1 beta mRNA level', 'IL - 6 mRNA level', 'TNF alpha mRNA level'. Ellipsis refers to the omission from a sentence or other construction of one or more words. There are numerous distinct types of ellipsis acknowledged in theoretical syntax.

In this paper, we describe the method for resolving noun phrases ellipsis in coordinate noun phrases. Our research is motivated by the need to reduce the costs of misunderstandings that can occur during NER. Without correct detection of omitted noun phrase, the syntactic analysis of the complete sentence fails in text mining.
In the rule-based approach, rules are constructed from a training dataset where an expert has assigned each sample. These rules will affect the performance. The quality of the rules is very important. When a domain is changed, it should be to create new rules. In the Machine Learning (ML) approach, it is necessary that the training data. Our research is motivated by the need to reduce the costs of misunderstandings that can occur during NER. We propose a method for resolving coordinate noun phrases with simple or complex ellipses using rules and dataset. And we present a method to automatically build dataset.

*Author for correspondence

The rest of the paper is structured as follows. In Section 2, we introduce the coordinate noun phrases problem in science literature. The buildings of a coordination noun phrases dataset are described in Section 3. Section 4 provides the description of the method for resolution of ellipses in coordinate noun phrases. Experimental results are reported in Section 5.

2. Related Works

Ambiguity is phenomenon inherent in natural language. It occurs everywhere in literature. In Natural Language (NL), ambiguity occurs when a sentence can be interpreted differently by different readers. In[1] discussed the problem of ambiguity in text and described the individual heuristics to predict innocuous ambiguity. In[2] developed the concept of an ambiguity threshold, which expresses the amount of variation between judgements that can be tolerated.

The reference[1,2] focused on coordination ambiguity, a common kind of structural ambiguity, highly prevalent in requirements documents. In the example E1, the coordination noun phrases 'cytokine and immune receptor' is the exemplification of the ambiguity pattern, where 'cytokine' and 'immune' are Far Conjunct (FC) and Near Conjunct (NC), and the noun 'receptor' is the attached Modifier (M). This can be resolved by two distinct bracketing:

1. [[Cytokine and immune] receptor].
2. [Cytokine] and [immune receptor].

(2) Said that (1) displays high attachment of the modifier, where M applies to both FC and NC, and displays love attachment, where M applies only to NC. The high attachments have semantic similarity between NC and FC. Machine learning approach is an established method for recognizing complex patterns automatically, based on empirical data, and learning of complex and nonlinear relation between data points.

The reference[3] defined the pre-conjunct and post-conjunct and suggested a rule based the parallelism to recognize a coordinating element. The reference[4] applied an unsupervised statistical model to resolve coordinate phrase attachment problem. Literature[5] evaluated to remove coordination ambiguity using word distributions in the corpus. Earlier studies usually focused mainly on conjunct identification. In[6] designed a post-processing rule to find the coordinate phrase. This method only identifies simple ellipses and could not recognized complex elliptical patterns. In[7] distinguished four categories of strings containing conjunctions: Named Internal Conjunction (NI), Named External Conjunction (NE), Right-Copy separator (RC), and Left-Copy Separator (LC). The categories of the conjunctions in the strings were assigned by a human annotator. It is difficult to classify according to this categories. And also machine learning approach for the coordination ambiguity problem in the newspaper domain[7]. They provided four types: named internal conjunction, named external conjunction, right-copy separator, and left-copy separator. Defined four categories of conjunction in named entity strings and used a supervised machine learning approach to conjunction disambiguation[7]. In[7] the candidate named entity string contains the conjunction and the conjunction serves to separate two distinct named entities than other categories. The reference[8] resolved the coordination ellipses in biomedical domain. And they used a machine learning approach based on Conditional Random Fields (CRFs) for recognizing elliptical entity expression. However, the method hard to identify nested entity mentions.

Machine learning approaches require an annotated corpus where named entities are labeled with their type and learn to decide if a word belongs to an entity of a particular type. Most of the well-known corpora for training and testing systems, also called gold standard. There are several linguistic resources. MUC-6 provided a corpus for English. There corpora cannot be obtained for free. MUC-7/MET-2 provided corpora for multilanguage. 2002 and 2003 editions of the Conference on Computational Natural Language Learning (CoNNL) provided corpora for multilanguage. Annotated texts that compose these gold standard corpora are informative texts obtained mostly from newswires. Wikipedia provides tools for their users to freely create and improve articles. GENIA[9] is an annotated research abstract corpus in molecular biology domain. The Colorado Richly Annotated Full Text Corpus[10] (CRAFT) is a manually annotated corpus consisting of 67 full-text biomedical journal articles. However, no corpus for resolution of coordinate noun phrases ellipses is available. As an alternative to overcome this problem, the automatic corpus construction approach called the silver standard corpus has come to the forefront in the field of NE research[11,12].

3. Dataset Construction

In this research, we need a dataset for the restoration of noun phrases ellipses in coordinate noun phrases. The

number of experts is necessary to maintain the objective to construct directly the dataset. This requires a lot of time to build a dictionary. It is difficult to generally apply to various domains (newswire, biomedical, disease etc). This is because the domain is changed to change the required dataset. We have built a domain independent dataset in order to overcome this limitation. If only the literature used in the domain, it is possible to automatically build dataset.

We have collected all the documents provided in PubMed by 2010 and extracted all NP present in the literature. The extracted NP is not to be included conjunction, preposition. If a NP appears more than once in the literature, we add the NP to the dataset for resolving NP ellipses. Table 1 shows the number of occurrences of the NP, within in the literatures. In the example E3, noun phrases are 'immunoregulatory gene expression'. This NP is counted the frequency in literature. The number of occurrence of the 'cytokine gene expression' is 2344 and 'viral gene expression' is 2962 in the literatures.

E3: Human cytomegalovirus binding to human monocytes induces immunoregulatory gene expression.

It is most difficult to distinguish between high modifier attachment and low modifier attachment in coordinate noun phrases. In this study, we constructed a new dataset using the NP in the literature in order to solve this problem. The process of building a dataset as follows:

- Collect noun phrases contained in the literature.
- Select the noun phrases that are likely to be the terminology.
- Normalize the noun phrase.

The selected NP is normalized through trimming and stemming. The trim removes the word with the parts of speech, such as 'DT' or 'PRP $' in the noun phrase. When the trim operation is complete, stemming operation is performed each word unit for each terminology. The stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected words to their word stem. This stemming can

Table 1. Frequency of noun phrases

Noun Phrases	CNT
immunoregulatory gene expression	5
cytokine gene expression	2344
viral gene expression	2962

Algorithm 1. Construct_noun phrases_dataset

```
input  : NP np
output: return is_ncnp
1  begin
2      is_cnp ← true;
3      is_nnp ← true;
4      have_core_np ← false;
5      is_ncnp ← false;
6      foreach word w of np do
7          pos ← GetPOS (w);
8          if pos ∈ {'IN','TO','WDT'} then
9              is_cnp ← false;
10         if pos ∈ {'CC'} then
11             is_nnp ← false;
12         if pos ∈ {'NN','NNS','NNP','NNPS','CD'} then
13             have_core_np ← true;
14     if is_cnp and is_nnp and have_core_np then
15         is_ncnp ← true;
```

improve the recall of a dictionary matching method. We used the stemming algorithms in Snowball. It has been used to find the stem for languages other than English and is a well-known stemmer in NLP. The normalized noun phrases through the trim and stemming are constructed as a dataset in order to resolve noun phrases ellipses. This dataset created automatically and used the terminology for resolution of noun phrases ellipses. And we constructed the automated dataset. These dataset are constantly evolving through new research. The dataset construction algorithm for noun phrases dataset as follows:

4. Method

We have implemented a system to detect noun phrases ellipses in coordinate noun phrases. Our system identifies ellipses in a textual input, and then resolves complex ellipses in coordinate noun phrases. It consists of four main functional process modules: Text Pre-processing Module, Noun Phrases Detection Module, Valid Noun Phrases Extraction Module, Noun Phrases Alignment Module, respectively.

In the Text Pre-processing Module, our system executes tokenization and normalization. In Second Module, our system is to detect the sentences that contain coordination structures, finds candidate noun phrases using the dataset, which is constructed in order to resolve noun phrases ellipses in Section 3. We used the scattered words matching algorithm[13]. Then the system determines candidate noun phrases group from the candidate noun phrases.

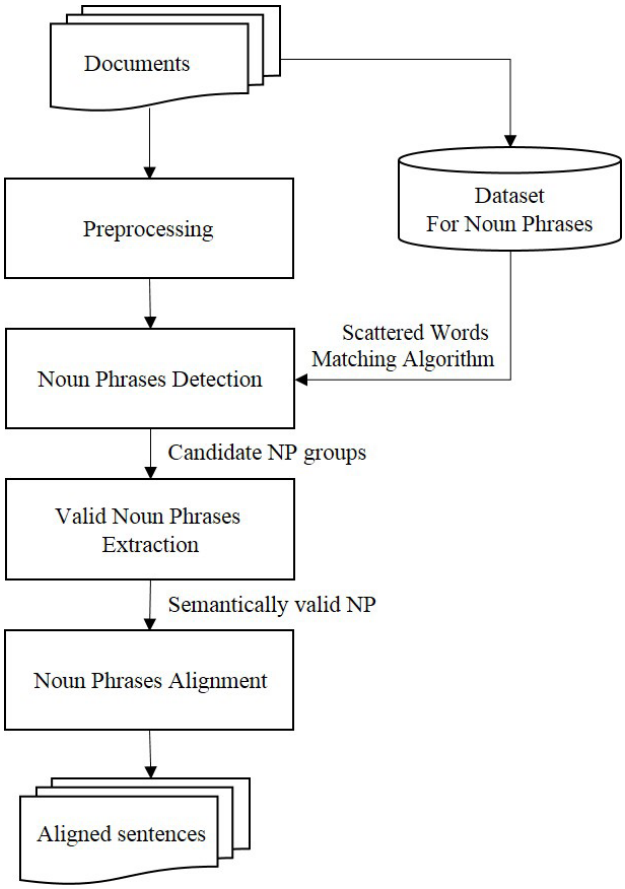

Figure 1. System architecture.

In Valid Noun Phrases Extraction Module, we determine semantically valid noun phrases from each candidate noun phrases group. Our system discards candidate noun phrases sets that do not satisfy the necessary condition of algebraic EMSEE[14]. The EMSEE usually contains one or more ellipses. It is converted into various different candidate algebraic EMSEEs, considering all the interpretation paths of the words. The algebraic operators can also represent relationships between phrases in sentences including coordinating conjunctions. This EMSEE is possible to extracts simple or complex ellipses using a construction dataset. Table 2 shows EMSEE and each letter is noun phrases.

Table 2. EMSEE in noun phrases

R1: (A * (B ⊕ C)
R2: (A - ⊕ B -) * A
R3: A * (B C * (- D ⊕ - E) ⊕ F * G * (- H ⊕ I))
R4: (A * B ⊕ C * (D ⊕ - E) ⊕ F * G) * H
R5: (A ⊕ B) * C * (D ⊕ E)
R6: ((A / B - ⊕ -) * C ⊕ D ⊕ E -) * F * G
R7: (A ⊕ B – C ⊕ D – E) * (F + ⊕ G -) * H

In Noun Phrases Alignment Module, we resolve ellipses of coordinate noun phrases. In this module, the noun phrase has not been recognized in the dataset. Because it has not been even once mentioned in PubMed literature. Noun phrase that does not appear in PubMed is unlikely to be used. For example 'T - helper 0 (th0) and 2 (th2) subset', noun phrases group follows as: T - helper 0 (th0) subset, T - helper 2 (th2) subset, T - helper 0 (th0), 2 (th2) subset. Because of that this NP never was used in PubMed literature. In this case, we apply low modifier attachment rule. Therefore our system extracts two noun phrases: T - helper 0 (th0) and 2 (th2) subset.

5. Results and Discussions

We used GENIA and CRAFT corpus to evaluate the performance of our system. The GENIA corpus has been frequently as the evaluation corpus NER system in the biomedical domain and 3,456 entity mentions from 1,596 coordinate noun phrases with ellipses. The CRAFT corpus is the linguistic annotation of 67 full-text biomedical publications. This corpus includes 1,036 entity mentions from 292 coordinate noun phrases with ellipses. We performed partial matching in these corpora.

Table 3 shows the performance evaluation results of identifying the non-elliptical noun phrases on the GENIA corpus. T1 consists of 80% of continuous entity mentions tagged in the corpus. T2 consists of 100% of continuous entity mentions tagged in the corpus. T3 is a full form which is created from the PubMed for evaluating the performance of noun phrases ellipsis. We create the T3, because of the absence of the gold standard corpus for the noun phrases ellipsis.

Table 4 shows the performance evaluation results of identifying the non-elliptical noun phrases on the CRAFT corpus. The experimental results of the proposed system using T1, T2, T3, T1+T3, T2+T3 are listed in Table 3, and 4. The system using T3 achieved high performance.

Table 3. Performance of the GENIA corpus

Type	*All*	*TP*	*FN*	*Recall (%)*
T1 (80%)	3,456	1,536	1,920	44.44
T2 (100%)	3,456	1,557	1,899	45.05
T3 (full form)	3,456	2,759	697	79.83
T1 + T3	3,456	2,769	687	80.12
T2 + T3	**3,456**	**2,770**	**686**	**80.15**

Table 4. Performance of the CRAFT corpus

Type	*All*	*TP*	*FN*	*Recall (%)*
T1 (80%)	1,036	734	302	70.85
T2 (100%)	1,036	744	292	71.81
T3 (full form)	1,036	856	180	82.63
T1 + T3	1,036	862	174	83.20
T2 + T3	1,036	862	174	83.20

6. Conclusions

In this study, we presented a method for resolving ellipses in coordinate noun phrases. Then, a dictionary-based approach was employed to identify non-ellipses noun phrases. We have built a domain independent dataset in order to overcome this limitation. If only the literature used in the domain, it is possible to automatically build dataset. We have collected all the documents provided in PubMed by 2010 and extracted all NP present in the literature. This method is applicable to a general-purpose in various fields. Our dataset effectively is used to distinguish between high and low modifier attachment. We reported on a set of experimental results to evaluate the performance of our approach. The results show that our system can efficiently resolve coordinate noun phrases. And we are sure that the method can resolve ellipses in various domains. It is necessary to extend it to a wider range of ellipses type, for example, to other types of structural ellipses like prepositional phrases, clause and sentence.

7. References

1. Chantree F, et al. Identifying nocuous ambiguities in natural language requirements. 14th IEEE International Conference in Requirements Engineering; 2006. p. 59–68.
2. Willis A, Chantree F, De Roeck A. Automatic identification of nocuous ambiguity. Research on Language and Computation. 2008; 6(3-4):355–74.
3. Agarwal R, Boggess L. A simple but useful approach to conjunct identification. Proceedings of the 30th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics; 1992. p. 15–21.
4. Goldberg M. An unsupervised model for statistically determining coordinate phrase attachment. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics; 1999. p. 610–4.
5. Chantree F, et al. Disambiguating coordinations using word distribution information. Proceedings of RANLP2005. 2005.
6. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002; 18(8):1124–32.
7. Dale R, Mazur P. Handling conjunctions in named entities. Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer; 2007. p. 131–42.
8. Buyko E, Tomanek K, Hahn U. Resolution of coordination ellipses in biological named entities using conditional random fields. 10th Conference of the Proceedings on Pacific Association for Computational Linguistics. PACLING 2007; 2007. p. 163–71.
9. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005; 6(1):57–71.
10. Bada M, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics. 2012; 13(1):161.
11. An J, Lee S, Lee GG. Automatic acquisition of named entity tagged corpus from world wide web. 41st Annual Meeting in Proceedings of Association for Computational Linguistics. 2003; 2:165–8.
12. Nothman J, et al. Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence. 2013; 194:151–75.
13. Chae J, et al. The partial matching method for effective recognizing HLA entities. J Korean Assoc Comput Educ. 2011; 14(2):83–94.
14. Chae J, et al. Identifying non-elliptical entity mentions in a coordinated NP with ellipses. J Biomed Informat. 2014; 47:139–52.