Toward Coreference Resolution in Scientific Domain

Aya Iwamoto, Hiroshi Noji, Hiroyuki Shindo, and Yuji Matsumoto

Graduate School of Information and Science
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, 630-0192, Japan
{iwamoto.aya.hr0, noji, shindo, matsu}@is.naist.jp

Abstract. Coreference resolution in scientific domain is necessary for further NLP tasks such as information extraction from scientific texts. For the traning data for implementing coreference resolution systems in scientific domains, the ACL anthology annotated data developed by Schäfer et al. [9] is useful. However, there are some difficulties in using this data, e.g., mistakes of part-of-speech tags because of unseen words (e.g., variables like 'a', 'x', etc.), and parsing errors often found in long sentences. Our contribution is two folds: (1) We modified annotated mention spans from entire noun phrases to minimum noun phrases that include dependency head (head NP), and (2) We added some information and re-formatted the data into the CoNLL shared task 2012 format.

Keywords: coreference resolution, scientific domain, resource

1 Introduction

Coreference resolution has been studied mostly in general domains (newspaper, conversation, etc.) [8], [5], [10]. The task is to find and cluster phrases (mentions) that refer to the same real world entities. In coreference resolution, we need to extract all candidate mentions (noun phrases, pronouns, named entities, ...), find anaphoric mentions and their antecedents, and then cluster mentions that refer to the same entity. Coreference resolution is a necessary task to improve other NLP tasks such as relation extraction. If we could enhance the performance of coreference resolution on scientific papers, it will improve the accerary of knowledge extraction from these documents. However, almost all coreference resolution systems are based on general domain data (such as OntoNotes [7] corpus). We use the ACL anthology annotated data [9], because we need a coreference resolution system for scientific documents.

Most coreference systems are comprised of two steps: (1) detecting mention candidates and then (2) performing coreference resolution on the detected candidates. Usually the candidates in the first step are extracted by some rules [1] such as extracting all noun phrases found by some parser. However we found that for the ACL anthology data such strategy causes problems. In particular we argue that the correct gold annotations in ACL anthology data hinders correct evaluation of performance. Below we describe this issue in detail in Section 2 and then present our proposed modifications to annotations in Section3.

2 Problems in the ACL anthology annotated data

In the ACL anthology annotated data, the mention spans are defined as the maximal noun phrases. Before doing coreference resolution, we need to extract all noun phrases as candidate mentions from parse trees. We parsed this data using Stanford CoreNLP [3], and found that the parser produced more errors than we had expected. There are two problems for the parser: Out-of-domain problem and sentence length problem. For the first problem, because many NLP tools including CoreNLP are usually tuned for general domains, these systems fail to parse scientific texts. For the second problem, sentences tend to be longer in scientific text than conventional data such as CoNLL shard task 2012 data based on OntoNotes corpus and this hurts parsing performance. Table 1 shows the relation between the average sentence length and the ratio of gold mentions covered by candidate mentions (that are identified by the parser). Note that the coverage of CoNLL test data is higher because it is based on the correct parses given by human annotators.

Table 1. Relation between average sentence length and gold mention coverage.

	ACL anthology	CoNLL 2012 shared task data		
		train	dev	test
Average Sententence Length	25.34	17.28	16.98	17.89
Gold Mention Coverage (%)	85.37	94.65	93.74	97.33

Therefore, we have difficulty in getting correct noun phrase boundary, which means we need to start with the predicted mention spans with low accuracy, which is too low compared with other data (e.g., CoNLL shared task 2012 data [6]). Figure 1 shows one example of parse error. in which the gold mention is *"pros and cons for each approach"* while the parper recognizes a number of noun phrases including *"pros and cons"* but not the exactly the same one as the gold mention.

This hurts the final evaluation of coreference: Even when the mention head word is correctly identified, if the detected mention span is wrong, correct pairs of antecedent and anaphor heads are regarded as wrong. Noun phrase spans are important but maximal NPs may not be necessary for further application. In order to avoid these problems, we modified annotated mention spans to minimal NPs, which include the dependency head word of the entire phrase.

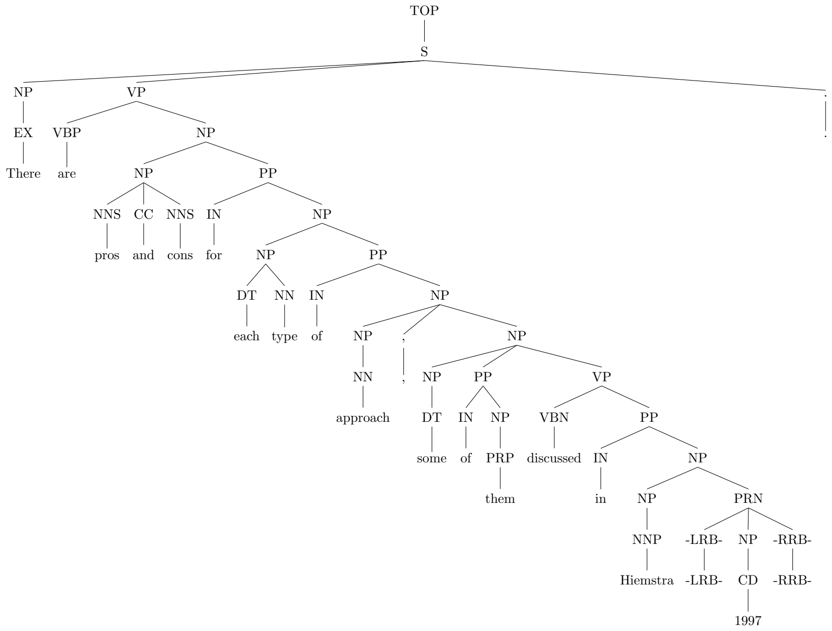

Fig. 1. Example of parse error. *"pros and cons for each type of approach"* is correct NP boundary.

3 Modifications to annotations

We changed the mention spans from maximal projections of NP to minimal head NPs. To extract the heads of mention spans, we first tried to analyze each mention using Stanford CoreNLP, but we found that often the parser parses it as a sentence (the root symbol is S, not NP). To parse mentions as noun phrases correctly, we trained Stanford parser [2] on the corpus comprised only of noun phrases, which we extracted from the parse trees in the Penn Treebank, then feed mentions to the trained parser. After that, the parsed noun phrase tree was passed to StanfordDependency [4] to determine head words.

Figure 2 shows an example of modified mention. Original mention span is (1) 'the determination algorithm described in (Mohri, 1997)' and (2) 'Mohri, 1997'. Mention (1) is a noun phrase and its head word is 'algorithm'. Because the minimal NP that contains 'algorithm' is 'the determination algorithm', we changed original mention (1) span to 'the determination algorithm'. For mention (2), because this is a citation, we didn't change the span.

For appositions, we separated them into two distinct mentions, and we parsed each phrase to determine the head noun phrase. Fig 3 shows another example mention which contains apposition. In this case, original annotated spans are (1) 'the transducer D' and (2) 'D'. They are both in the same coreference set,

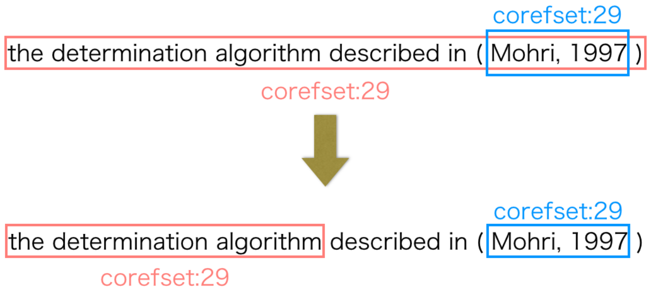
corefset:29
the determination algorithm described in (Mohri, 1997)
corefset:29
corefset:29
the determination algorithm described in (Mohri, 1997)
corefset:29

Fig. 2. Example of modified mention.

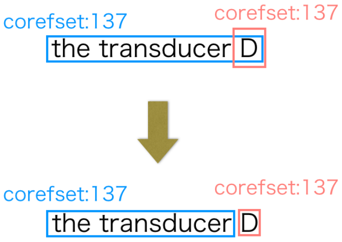

Fig. 3. Example of modified apposition mention.

so we regarded these two mentions are appositions. We changed mention (1) to 'the transducer'.

In case a whole coordinate structure is referred to by other mentions, we did not change the annotated span of coordinate structures. For example, a mention like 'violate match, exact match and inside match' has coordinate structure (see Fig 4). This mention is referring antecedent 'three cases of structure matching'. If we change the span to head NP 'violate match', because head NP of coordinate phrase is the first element of the phrase, it will mean another entity ("violate match, exact match and inside match") and has exactly same span with other mention.

4 Result and conclusion

Table 2. statistics before & after the change of mention spans.

	before	after
# of documents	266	262
# of mentions	83775	82027
coverage	85.37	90.04

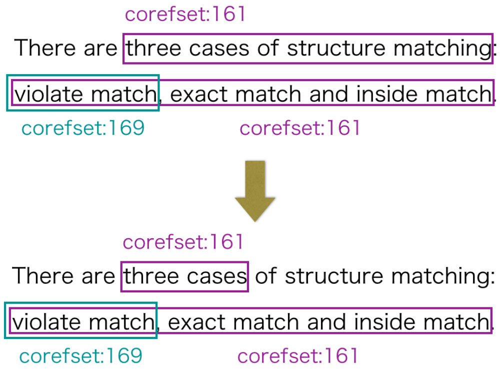
corefset:161
There are three cases of structure matching:
violate match, exact match and inside match.
corefset:169
corefset:161
corefset:161
There are three cases of structure matching:
violate match, exact match and inside match.
corefset:169
corefset:161

Fig. 4. Example of coordinate structure.

After modifying the annotations to head NPs, the coverage of mention detection increased from 85.37% to 90.04%. The detail of this change is shown in Table 2. The number of documents are decresed because of the inconsistency in the data. In four documents, there are mentions which refers different entities in one mention.

By doing this, we can evaluate the performance of coreference system based on mention's head NPs. For the future work, we have to evaluate the automatic modification of annotation. Because the NP parser is based on general domain, we have to check the accuracy of head word extraction.

Acknowledgement

This work was supported by CREST, JST.

References

1. Durrett, G., Klein, D.: Easy victories and uphill battles in coreference resolution. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1971–1982. Association for Computational Linguistics, Seattle, Washington, USA (October 2013), http://www.aclweb.org/anthology/D13-1203
2. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. pp. 423–430. Association for Computational Linguistics, Sapporo, Japan (July 2003), http://www.aclweb.org/anthology/P03-1054
3. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010

4. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation(LREC'06). pp. 449–454. Association for Computational Linguistics (2006)
5. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. pp. 104–111. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (July 2002), http://www.aclweb.org/anthology/P02-1014
6. Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y.: Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In: Joint Conference on EMNLP and CoNLL - Shared Task. pp. 1–40. Association for Computational Linguistics, Jeju Island, Korea (July 2012), http://www.aclweb.org/anthology/W12-4501
7. Pradhan, S.S., Xue, N.: Ontonotes: The 90% solution. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts. pp. 11–12. Association for Computational Linguistics, Boulder, Colorado (May 2009), http://www.aclweb.org/anthology/N/N09/N09-4006
8. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.: A multi-pass sieve for coreference resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 492–501. Association for Computational Linguistics, Cambridge, MA (October 2010), http://www.aclweb.org/anthology/D10-1048
9. Schäfer, U., Spurk, C., Steffen, J.: A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In: Proceedings of COLING 2012: Posters. pp. 1059–1070. The COLING 2012 Organizing Committee, Mumbai, India (December 2012), http://www.aclweb.org/anthology/C12-2103
10. Soon, W.M., Ng, H.T., Chung, D., Lim, Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27, 521–544 (2001)