Corpus for Coreference Resolution on Scientific Papers

Panot Chaimongkol, Akiko Aizawa, Yuka Tateisi

The University of Tokyo, National Institute of Informatics
Tokyo, Japan
melsk125@nii.ac.jp, aizawa@nii.ac.jp, yucca@nii.ac.jp

Abstract

The ever-growing number of published scientific papers prompts the need for automatic knowledge extraction to help scientists keep up with the state-of-the-art in their respective fields. To construct a good knowledge extraction system, annotated corpora in the scientific domain are required to train machine learning models. As described in this paper, we have constructed an annotated corpus for coreference resolution in multiple scientific domains, based on an existing corpus. We have modified the annotation scheme from Message Understanding Conference to better suit scientific texts. Then we applied that to the corpus. The annotated corpus is then compared with corpora in general domains in terms of distribution of resolution classes and performance of the Stanford Dcoref coreference resolver. Through these comparisons, we have demonstrated quantitatively that our manually annotated corpus differs from a general-domain corpus, which suggests deep differences between general-domain texts and scientific texts and which shows that different approaches can be made to tackle coreference resolution for general texts and scientific texts.

Keywords: coreference resolution, annotated corpus, scientific domain

1. Introduction

The number of published scientific papers has increased at an ever-growing rate for decades (Larsen and von Ins, 2010). That growth therefore prompts the need for automatic knowledge extraction from these publications to help scientists understand the current state of knowledge in a given field or across fields. Coreference resolution is an automatic knowledge extraction task. It was designed to identify all the mentions in a text or a collection of texts that refer to the same entity.

There are a number of shared tasks and corresponding corpora for the task of coreference resolution in a general domain, such as Message Understanding Conference (MUC) conferences (MUC, 1995) (Chinchor, 1998), and Automatic Content Evaluation (ACE) Program conferences (Doddington et al., 2004). These corpora mainly include news article and newswire texts.

To serve as training and evaluating data for scientific coreference resolvers, we therefore develop a scientific domain coreference corpus. As described in this paper, we present the coreference-annotated corpus that consists of the abstracts of 284 articles from four different scientific fields, taken from the corpus for the shared task SemEval-2010 Task 5 (Kim et al., 2010). This shared task is a keyphrase extraction task. The original data also contain keyphrases for each article.

In section 2, we describe where we get and how we annotate the data. In section 3, we present statistics of the corpus in a similar fashion to the manner used in an earlier report (Stoyanov et al., 2009). We give results obtained from application of an existing coreference resolver on our data in section 4. A discussion of our work is presented in section 5. Then we present conclusions in section 6.

2. Resource Creation

In this section, we describe the source of our data and our method in annotating the corpus. The corpus can be downloaded at http://github.com/melsk125/SciCorefCorpus.

2.1. Source Data

We choose the corpus from SemEval-2010 Task 5 (Kim et al., 2010) as our source data. The corpus includes articles from different scientific fields and is already annotated for another task: automatic keyphrase extraction.

The corpus contains 284 scientific articles from the ACM Digital Library of four different fields. The articles are grouped according to the digital library classification into the following classes: C2.4 (Distributed Systems), H3.3 (Information Search and Retrieval), I2.11 (Distributed Artificial Intelligence – Multiagent Systems), and J4 (Social and Behavioral Sciences – Economics). The articles are, originally in the corpus, divided into three sets: a trial set (40 articles), a training set (144 articles), and a test set (100 articles). The training set is intended for use in training a machine learning model. The trial set (or development set) is intended for use as a test set while developing the model.

We used only the abstracts of the articles because of time and resource limitations. The articles are apparently read automatically by an OCR system. Therefore, noise and mid-sentence newlines are present in the data. We cleaned up the text by viewing the associated PDF files from the ACM Digital Library.

2.2. Manual Annotation

We use brat (Stenetorp et al., 2012) as the annotation tool for our task. The distributed data are therefore in the standoff format, which is the default format for brat. The annotation instructions we gave to the annotator are an extension of the annotation guideline of MUC-6 task. However, our annotation guideline differs from that of MUC-6 task in the following ways.

- *Named entities.* Date, time currency expression, and percentage are numerical expressions that are defined

as named entities in the MUC-6 task. We also regard several other numerical expressions as markable. Numerical quantities such as the value of a variable are regarded as markable; they are intended to be linked as coreferent with the variable or their description. Moreover, mathematical expressions that are not propositions, i.e. expressions that do not hold a truth value such as variable, functions, and their applications, are markable. We consider that these extended named entities appear more often in scientific articles. Therefore, they are useful when they are marked in coreference chains.

- *Relative pronouns.* We require the annotator to mark relative pronouns such as *which* and *that*.
- *Conjoined noun phrase.* In our scheme, the annotator is instructed to mark any noun phrase, conjoined or not, whenever possible. The MUC-6 annotation guideline considers noun phrases with two or more head tokens non-markable because annotators cannot identify their unique contiguous head substring. Because our annotation scheme requires no annotator to mark the head substring of markables, the restriction can be relaxed.

3. Corpus Statistics

Our corpus includes 4,228 mentions and 1,362 coreference chains. The average length of a chain is 3.1 mentions.

To characterize our corpus in contrast to corpora in general domains, we follow the resolution class analysis described in an earlier report (Stoyanov et al., 2009). They define nine resolution classes. Classes PN-e, PN-p, and PN-n are resolution classes of proper names. PN-e is a proper name with at least one preceding markable in its coreference chain that exactly matches it. Actually, PN-p is a proper name with at least one preceding markable in its coreference chain that has some content words in common. In fact, PN-n is a proper name that has no string match with preceding mentions in its coreference chain. Classes CN-e, CN-p, and CN-n are resolution classes of common noun phrases that are analogous to PN classes. The last three classes are pronouns: 1+2Pr is a first or second person pronoun, G3Pr is a gendered third person pronoun, and U3Pr is an ungendered third person pronoun.

The distribution of resolution classes in our corpus, together with that of two MUC corpora and four ACE corpora are given in Table 1. To confirm that general-domain texts are similar, we calculate the Pearson correlation coefficient between the distribution of each corpus and the sum of the distribution of five other general-domain corpora. Furthermore, to answer the question of whether scientific texts differ from general-domain texts, we calculated the correlation between the distribution of our corpus and the sum of six general-domain corpora. The correlations are also given in the Table 1.

Results showed that general-domain corpora have high correlation with the average: the minimum is 0.60 and maximum is 0.88. However, our manually annotated scientific-text corpus has -0.10 correlation with the average. Results show that the resolution classes in our manually annotated corpus are distributed very differently from those of the general domain corpus. Therefore, we conclude that general domain texts and scientific texts differ markedly from the coreference perspective.

4. Performance on Existing Coreference Resolver System

To compare the coreference resolution result of our corpus with those of other corpora, we have tested our corpus on an existing coreference system: Stanford CoreNLP Coreference Resolver (Lee et al., 2013). Table 2 shows the performance of the resolver on MUC-6 corpus, ACE2004 corpus, and our corpus. The score on MUC-6 and ACE2004 corpora is the reported score on the tool web-site. It is the score given when the resolver is given gold mentions of the text. For comparison, we apply the resolver to our corpus on the settings both using and not using gold mentions. MUC (Vilain et al., 1995) and B-CUBED (Bagga and Baldwin, 1998) scoring schemes are applied to measure the precision, recall, and F1 score for each corpus.

The F1 score on our corpus in both gold-mention and detected-mention settings is lower than other corpora. For gold-mention setting, the resolver performs better on our corpus in terms of precision than most of other corpora, but worse in terms of recall. The score for MUC-6 test set, which has an annotation scheme that most resembles that of our corpus, dominates the score on our corpus. Moreover, when we do not use gold mention of our corpus, the MUC score drops dramatically. Results show a greater than 10 score drop in B-CUBED precision and small improvement in B-CUBED recall.

5. Discussion

Many coreference corpora exist in specific scientific domains. For instance, (Cohen et al., 2010) presents a corpus of 97 full-text journal articles in the biomedical domain and (Schäfer et al., 2012) is a corpus fully annotated 266 scientific articles from the ACL Anthology, a collection of articles from the field of computational linguistics.

However, no coreference corpus has been reported for use in multiple scientific fields. Different scientific fields arguably have different sets of vocabulary and also writing styles. To build a coreference system based on only one scientific field, whether heuristically or with machine learning method, might produce a resolver that is specific only to a single scientific field and which is not applicable to other fields.

Moreover, we have quantitatively compared our corpus, which comprises scientific texts, with a general-domain corpus. This can provide insight into the difference between two types of texts, which previously described works have not done.

Furthermore, our corpus is built upon another annotated corpus. Consequently, multiple annotations are freely available for these particular data. This point can be useful for researchers who want to use information of multiple types for the same text collection.

	MUC-6	MUC-7	ACE2	ACE03	ACE04	ACE05	Sum	Our corpus
PN-e	273	249	346	435	267	373	1943	94
PN-p	157	79	116	178	194	125	849	69
PN-n	18	18	85	79	66	89	355	34
CN-e	292	276	84	186	165	134	1137	432
CN-p	229	239	147	168	147	147	1077	784
CN-e	194	148	152	148	266	121	1029	399
1+2P	48	65	122	76	158	51	520	685
G3Pr	160	50	181	237	246	69	943	17
U3Pr	175	142	163	122	153	91	846	352
Sum	154	1266	1396	1629	1662	1200	8699	2866
Correl	0.77	0.64	0.76	0.88	0.60	0.84		-0.10

Table 1: Frequency for each resolution class in each corpus and the correlation of the corpus against the sum of general domain corpora.

	MUC			B-CUBED		
	Precision	Recall	F1	Precision	Recall	F1
ACE2004 dev	86.0	75.5	80.4	89.3	76.5	82.4
ACE2004 test	82.7	70.2	75.9	88.7	74.5	81.0
ACE2004 nwire	84.6	75.1	79.6	87.3	74.1	80.2
MUC-6 test	90.6	69.1	78.4	90.6	63.1	74.4
Our corpus (gold mention)	86.8	53.4	66.1	91.5	62.6	74.3
Our corpus (detected mention)	49.8	38.2	43.2	77.0	64.3	70.1

Table 2: Result of Stanford CoreNLP Coreference Resolver on our corpus and existing general domain corpora.

6. Conclusion

We have annotated a coreference resolution corpus in multiple scientific domains, and have furthered the process of annotation. We have also developed annotation guidelines for coreference in scientific texts that can be viewed as an extension of MUC annotation guidelines. Our corpus consists of different scientific fields. It is shown to be different quantitatively from general domain corpora. It is challenging for an existing coreference resolver. We hope that our corpus is useful for the scientific community. We welcome improvements of our corpus by the community.

7. References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Nancy A. Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of the 7th Conference on Message Understanding*, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence E. Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of the 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM-2010)*, pages 37–41.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) program - tasks, data, and evaluation. In *LREC*. European Language Resources Association.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Los Angeles, California. Association for Computational Linguistics.

Peder Olesen Larsen and Markus von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

1995. Appendix D: Coreference task definition (v2.3). In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 335–344, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ulrich Schäfer, Christian Spurk, and Jörg Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In *Proceedings of COLING 2012: Posters*, pages 1059–1070, Mumbai, India, December. The COLING 2012 Organizing Committee.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012.

brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.