Utilising Technical Term Extraction in Coreference Resolution on General Academic Domains

Panot Chaimongkol Akiko Aizawa
Department of Computer Science, University of Tokyo, Japan
National Institute of Informatics

{melsk125, aizawa}@.nii.ac.jp

1 Introduction

Coreference resolution, the task of identifying mentions that refer to the same entity, is an important task for natural language understanding such as question answering, summarising, and information extraction (IE). It is a task that looks simple for human beings who intuitively and repeatedly resolve coreferring mentions every time we encounter anaphoric expressions or rephrases, yet it is not trivial for automated systems due to the ambiguous nature of natural languages.

Coreference resolution has a long history with many proposed techniques [2, 9, 10] and datasets [4, 6, 13, 14]. Most of them focus on texts in general domain, such as news articles, allowing automated systems to aggregate information from digital documents. However, the approaches in academic domains are very limited. To the best of our knowledge, the only field that has been widely focused is biomedicine. There is also a coreference resolution corpus in computational linguistics with documents from ACL Anthology **??**. However, there has been only a few or no attempts in tackling the problem in general academic domains.

After preliminary investigation on scientific abstracts from several domains, we noticed that technical terms, lexical units that are used in a more or less specialised way in a domain [7], are promising candidates for coreference mentions in academic domains, since technical terms can be considered as main participants in academic writings. We thus focus on improving coreference resolution on general academic domains, utilising extracted technical terms.

We have created datasets for technical term extraction and coreference resolution based on a corpus of multiple academic domains. The datasets are used in training and testing our term extraction system, and evaluating our proposed methods of integrating term extraction result into an existing coreference resolution system, namely Stanford's Dcoref coreference resolver [9, 10].

In section 2, we provide information about previous works in technical term extraction and coreference resolution that are related to ours. We describe the characteristics of out dataset in section 3. The details of our technical term extraction unit are presented in section 4. Section 5 shows our methods to integrate the result of term extraction into mention detection module. We conclude our work in section 6

2 Related works

2.1 Term extraction

The *C-value/NC-value* method [5] is a commonly used statistical method for technical term extraction. The method introduces the concept of *termhood*, a statistical characteristic of candidate phrases, and uses it in a combination with context information. [5] and other related studies reported that the method performed better than simple dictionary and frequency counting, especially for nested candidates.

Recently, Dirichlet Process segmentation (DP-seg), an unsupervised model, has been used for identifying correct spans in index term and keyphrase extraction [11]. DP-seg reportedly outperforms the conventional C-value/NC-value.

Evaluation of term or keyphrase extraction is hindered with the lack of extensively annotated corpus. Both [5] and [11] argued that constructing such corpora is costly, and had resorted to evaluate their works solely with precision rather than the commonly used F1 measure.

2.2 Coreference resolution

Most of the freely available corpora for coreference resolution are in general domain. MUC-6 [6] is one of the first shared tasks aiming at resolving coreferences. The shared task also introduced the MUC scoring algorithm, which has become one of the standard scoring algorithms up to now.

Other noteworthy shared task of coreference resolution in general domains are ACE2004 [4] and

CoNLL-2011 [13]. Both MUC-6 and ACE2004 provide a standard for annotation scheme that was later adopted in other corpora. CoNLL-2011 is one of the latest shared tasks on coreference resolution. It had 18 participating teams with a variety of methods, among which the simplistic rule-based method performed the best [10]. The rule-based method, Dcoref, is reportedly inspired by the idea of resolving coreference starting with a set of very high-precision contraints [2]. We chose Dcoref as the basis of our work because of two reasons: it is one of the best-performing systems that is freely available, and it is a modular, easily expandable package.

3 Dataset

We use the corpus from SemEval-2010 Task 5 [8], because the corpus is comprised of articles from multiple scientific fields, which is different from corpora solely in biomedical domain [3] or computational linguistics [14].

The corpus contains 284 scientific articles from the ACM Digital Library of 4 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 3 sets: trial set (40 articles), training set (144 articles), and test set (100 articles).

Since the corpus was originally designed for automatic keyphrase extraction, we thus annotated the corpus for term extraction and coreference resolution tasks. We used only on the abstracts of the articles, due to time and resource limitations.

For term extraction, all the technical terms in the target corpus should be annotated. We adopt the definition of technical terms from [7]: [Technical terms are] lexical units used in a more or less specialised way in a domain. Since we have chosen to use CRFs as the extraction model, the identified terms are not to be overlapped. We thus require the annotator to identify the maximal spans that are considered terms. Moreover, the spans should not include articles, such as *a*, *an*, and *the*.

The statistics for the technical term annotation is given in Table 1.

	Train	Dev	Test
Number of documents	104	40	100
Number of term spans	1,725	761	1,857
Average number of term spans in a document	16.59	19.03	18.57

Table 1: Statistics for term extraction corpus

For coreference resolution, we follow the definition of the MUC-6 coreference task [1] as the base of our annotation scheme. However, we apply the following extensions in order to suit coreference data in scientific context.

- *Named entities.* Date, time currency expression, and percentage are numerical expressions that are defined as named entities in the MUC-6 task. We also consider numerical quantities such as value of a variable and mathematical expressions that are not proposition as markable. We consider these extended named entities appear more often in scientific articles and thus useful when they are marked in coreference chains.
- *Relative pronouns.* We require the annotator to mark relative pronouns, such as *which* or *that.*
- *Conjoined noun phrase.* In our scheme, the annotator is instructed to mark any noun phrases, conjoined or not, whenever possible. MUC-6 annotation guideline considers noun phrases with two or more head tokens non-markable, because annotators cannot identify their unique contiguous head substring. Since our annotation scheme, on the other hand, does not require the annotator to mark head substring of markables, the restriction can be relaxed.

The resulting annotated corpus contains 4,228 mentions and 1,362 coreference chains, while the average length of the chains is 3.1 mentions.

4 Technical term extraction

For technical term extraction, we formulate the task as sequential labeling and use conditional random field (CRF) as the model. We adopt the BIO-labeling scheme and train our model with the training data.

We use 2 feature sets to characterise tokens, context features (C) and orthographic features (O).

- *Context features* consists of word surface and part-of-speech tag up to 5-token windows centered on the token in question.
- *Orthographic features* consists of 25 features to capture the orthographic characteristic of the token.

We use NLTK [1] to perform sentence and word tokenisaion and part-of-speech tagging. For CRF implementation, we use CRFsuite [12].

The performance of our system is given in Table 2. We also provide the performance score from dictionary method as a baseline. In the dictionary method the system collects all the technical terms in training set and marks only the exactly matching spans in target texts.

[1] http://nltk.org/

	Precision	Recall	F1
TermDict	49.02	33.11	39.52
CRF C	64.07	41.00	50.00
CRF CO	**69.75**	**54.33**	**61.21**

Table 2: Performance of term extraction system

5 Coreference resolution

5.1 Dcoref coreference resolver

We aim to improve the existing Dcoref coreference resolver developed by a research group at Stanford University [9, 10]. The two main reasons are that the resolver is a modular and open-sourced Java package, and that the resolver performs very well despite of its simplicity.

Decoref is a rule-based coreference resolver, and is composed of two main phases, *mention detection* and *mention linking*.

Mention detection module first selects all noun phrases, pronouns, and named entities as candidates and then exclude some candidates corresponding to some specific rules. The resulting candidates are expected to have high recall.

Mention linking module is comprised of smaller modules called *sieves*. Each sieve links two mentions that correspond to its condition. The sieves are sorted in the descending order of precision to minimize the number of incorrect links.

The system was the best at the CoNLL-2011 contest, but with our SciCorefCorpus, the performance dropped, as seen in Table 3. The performance of the mention detection module is also given in the last row.

	Precision	Recall	F1
Pairwise	57.2	41.7	48.2
MUC	55.1	42.6	48.0
B-Cubed	52.7	39.8	45.4
Mention	27.0	75.4	39.8

Table 3: Performance of Dcoref on SciCorefCorpus test set

5.2 Proposed method

In this paper, we focus on improving the precision of mention detection module, because the precision of the mention detection module is at 27.0 percentage points, as shown in Table 3, leaving a huge room for improvement.

To improve the precision of mention detection module, we combine the detected mentions with the extracted term using *CRF-CO* model as we described in Section 4. The idea is of improvement is that we focus mainly on technical terms as mentions to be resolved. We investigate the following four combinations.

1. Use extracted terms.
2. Use detected mentions that contain at least a single extracted term as substrings.
3. Use detected mentions overlap at least a single extracted term.
4. Use the union of detected mentions and extracted terms.

Since pronouns constitute a large portion in coreference chains, we also add pronouns into the extracted set before combining with the detected mentions.

The result of the combinations with pronouns are shown in Table 4. The mention detection performance is also given in the last row. The boldface numbers indicate the top performer in the corresponding scoring measure.

Combination		Precision	Recall	F1
Comb. 1	Pairwise	60.8	32.2	42.1
	MUC	59.4	28.4	38.4
	B-Cubed	56.3	24.9	34.5
	Mention	36.1	35.5	35.8
Comb. 2	Pairwise	71.7	35.4	**47.4**
	MUC	**69.5**	33.4	**45.1**
	B-Cubed	66.9	29.5	41.0
	Mention	37.2	50.6	**42.9**
Comb. 3	Pairwise	**72.0**	35.0	47.1
	MUC	70.0	32.7	44.5
	B-Cubed	**67.1**	28.5	40.0
	Mention	**38.6**	47.9	42.7
Comb. 4	Pairwise	42.4	**41.3**	41.8
	MUC	45.7	**42.2**	43.9
	B-Cubed	43.1	**39.4**	**41.2**
	Mention	24.0	**78.6**	36.7

Table 4: Coreference resolution performance with extracted terms and detected pronouns

In Table 3 and 4, we see that combinations 1–3 yield higher precision for mention detection but lower recall than the original system, while combination 4 gets higher recall and lower precision. Moreover, combinations 2 and 3 also perform better than the original system in term of F1 score for mention detection. Combination 2, with the higher F1 score in mention detection, suffers nearly 25 points drop in recall, but the 10 points gain in precision pushes F1 score up by 3 points. Combination 4, which is a union of extracted terms and detected mentions, gains higher recall and lower precision as expected.

For coreference scores, combinations 1–3 show higher precision scores, compared to the original system, but the recall measures are lower. Meanwhile, combination 4 does not gain a higher recall scores despite the higher mention detection recall.

The are cases where dropping unrelated candidates improves the coreference result. In these cases, the sieves correctly link remaining links, where they make mistakes when the recall of mention detection is higher in the original system. This partially contributes to the precision measures. Meanwhile, we see that the number of coreference chains and linked mentions are much lower for combinations 1–3. This also contributes to the precision measures but largely damages recall scores.

6 Conclusion

We proposed methods to incorporate extracted technical terms into mention detection module of the Dcoref coreference resolution system. Our technical term extraction system used an in-house annotated dataset to train CRF models and was evaluated with the corpus. The integrated mention candidates were then pushed through sieves in the Dcoref system. Finally, we evaluated the effectiveness of our method with an in-house dataset, and found that our method performed better in term of precision.

There are many limitations in our work. Our method eliminates all mention candidates that do not overlap with extracted terms. We should consider relaxing this restriction to allow some non-overlapping candidates to be included if they satisfy some conditions, such as C-value/NC-value. Furthermore, we want to evaluate our method with other corpora, such as a coreference corpus from ACL Anthology [14].

References

[1] Appendix D: Coreference task definition (v2.3). *Proceedings of the 6th Conference on Message Understanding, MUC6 '95.* pp. 335–344, 1995

[2] Baldwin, B. CogNIAC: High precision coreference with limited knowledge and linguistic resources. *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, ANARESOLUTION '97.* pp. 38–45, 1997.

[3] Cohen, K. B., Lanfranchi, A., Corvey, W,. Baumgartner Jr., W. A., Roeder, C., Orgen P. V., Palmer, M., Hunter, L. Annotation of all coreference in biomedical text: Guideline selection and adaptation *Proceedings of the 2nd Workshop on Build- ing and Evaluating Resources for Biomedical Text Mining (BioTxtM-2010)* pp. 37–41, 2010.

[4] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ranshaw, L. A., Strassel, S., Weischedel, R. M. Automatic Content Extraction (ACE) program - tasks, data, and evaluation. *LREC* 2004.

[5] Frantzi, K., Ananiadou, S., and Mima, H. Automatic recognition of multi-word terms:. the c-value/nc-value method. *Int. J. Digital Libraries*, 3:115—130, 2000.

[6] Grishman, R., Sundheim, B. Message understanding conference-6: A brief history. *Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING '96* pp. 466–471, 1996.

[7] Kageura, K. *The Quantitative Analysis of the Dynamics and Structure of Terminologies* vol.15, 2012.

[8] Kim, S. N., Medelyan, O., Kan, M., Baldwin, T. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation.* pp. 21–26, 2010.

[9] Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics.* 39(4):885–916, 2013.

[10] Lee, H., Peirsman, Y., Chang. A., Chambers, N., Surdeanu, M., Jurafsky, D. Stanford ̓s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task.* pp. 28–34, 2011.

[11] Newman, D., Koilada, N., Lau, J. H., and Baldwin, T. Bayesian Text Segmentation for Index Term Identification and Keyphrase Extraction. *Proc. COLING 2012.* pp. 2077-2092, 2012.

[12] Okazaki, N. *CRFsuite: a fast implementation of Conditional Random Fields (CRFs)*, 2007.

[13] Pradhan, S., Ranshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N. CoNLL-2011 shared task: Modeling un restricted coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* pp. 1–27, 2011.

[14] Schäfer, U., Spurk, C., Steffen, J. A fully coreference-annotated corpus of scholarly papers from the ACL anthology. *Proceedings of COLING 2012: Posters* pp. 1059–1070, 2012.