Utilizing LDA Clustering for Technical Term Extraction

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

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

1 Introduction

Similar to named entities in general texts, technical terms are critical in scientific writings. They appear in nearly all parts of the document structure, and as such, identifying technical terms in a scientific writing is arguably the first step to analyze the semantic content. We thus focus on technical term extraction in this paper, a task of recognizing technical terms from scientific writings.

The motivations behind includes a possibility of assisting researchers in exploring vast amount of scientific publications in their research activities. Technical term extraction would benefit advanced information extraction (IE) systems and eventually lead to computer-aided inference of knowledge contained in these publications.

In this paper, we propose a simple unsupervised method for technical term extraction by combining Conditional Random Field (CRF) with extra features obtained from a Latent Dirichlet Allocation (LDA) clustering model. We found that such a combination is promising to improve the performance of technical term extraction.

In section 3, the details of our proposed method are described, including the problem formulation and how we incorporate LDA clustering information in the feature set. Section 4 describes the experimental design and settings to investigate the performance of our method. Results of the experiments are provided in section 5, followed by conclusions and discussions for further improvement in section 6.

2 Related works

2.1 Term extraction

The most commonly used method in technical term extraction is C-value/NC-value proposed in [4]. It is a combination of *termhood*, statistical characteristics of candidate phrases, and context information. In [4] and other related studies, it is reported that C-value/NC-value performed better than mere frequency counting especially on candidates which only appeared as nested.

Recently, [7] proposed an unsupervised model called Dirichlet Process segmentation (DP-seg) for identifying correct spans in index term and keyphrase extraction. The experimental results showed that the proposed DP-seg model outperformed the conventional C-value/NC-value method.

As for the evaluation of term or keyphrase extraction, both [4] and [7] argued that constructing extensively annotated corpora is too costful. Without widely available datasets that enumerate all the terms contained in the text, most conventional evaluations rely solely on the precision score rather than the F1 score commonly used in NER.

2.2 Named entity recognition

One of the general approaches to NER is to pose the problem as sequential labeling and use supervised machine learning methods to build the model.

One of the general approaches to NER is to pose the problem as sequential labeling and use supervised machine learning methods to build the model. CRF, a probabilistic model, was proposed in [5] together with a performance experiment in part-of-speech tagging. It was shown in [5] that CRF outperformed Hidden Markov Model (HMM) and Maximum Entropy Maykov Model (MEMM).

Since CRF is trained by supervised learning algorithms, the most important key to the success of the model is the choice of feature sets. Apart from conventional features embedding contextual and word shape information, [6] showed that extra clustered resources help dsicriminative classifiers such as CRF to perform better in named-extity recognition (NER). The paper applied K-Means algorithm to cluster phrases and used the cluster and its information as features of token. The systems utilizing cluster information reportedly outperformed the baseline CRF system.

3 Proposed method

3.1 Overview of the proposed method

Since the number of annotated corpora in scientific domain is very limited, it is necessary to utilize extra resources to provide the system with background knowledge. In this paper, we thus consider a unsupervised framework similar to [6] that has been successfully applied to NER.

Our framework utilizes Latent Dirichlet allocation (LDA) model which is an unsupervised, generative model designed for, but not limited to, capturing topics of given documents [2]. We consider that LDA is more appropriate to cluster scientific papers than K-Means algorithm used in [6] because it can capture the different terminologies in different academic fields. Since each scientific paper is arguably written in a specific field of knowledge, such as chemistry or physics, the sets of technical terms for different fields might help improving the system. The detailed explanation of how we apply LDA model is given in section 3.3.

3.2 Term extraction using CRF

We pose Technical Term Extraction as sequential labeling problem, i.e. putting labels on linearly ordered data such as string of words or time series, while assuming relations between two consecutive items.

We formulate the problem in the same way as our previous work [3]. The input for our system is a text $\mathcal{T}$ of length T tokens, which is tokenized as $tok_1, \ldots, tok_T$. The output is a length T string of tags in BIO tagset, $tag_1, \ldots, tag_T$, where $tag_i = B$ when tok_i is the beginning of a technical term, $tag_i = I$ when tok_i is in a term but not the beginning, and $tag_i = O$ otherwise.

In situations when fully annotated gold-standard corpus is not available, a supervised method cannot be applied. However, some corpora of scientific papers come together with author-assigned keywords and it is possible to assume that such keywords are technical terms. We thus automatically tag the corpus with keywords and treat it as the training set for CRF supervised model, which would also tag similar words as technical terms.

3.3 LDA-based features

We propose two clustering schemes, namely *document clustering* and *noun phrase clustering*, and include information about the cluster as feature in CRF model.

3.3.1 Document clustering

In document clustering scheme, each document is clustered according to its tokens. Each token appeared in a document is counted and the frequencies of tokens represent the document. This clustering scheme is the typical application of LDA and is aimed for classifying scientific documents into the corresponding fields.

The LDA model is pre-trained in an unsupervised manner with a training set. A queried document is then classified into a cluster according to the trained model. Every tokens in the document include features indicating the cluster.

3.3.2 Noun phrase clustering

In noun phrase clustering scheme, each noun phrase is clustered according to verb frames with which it co-occurs. Each verb frame co-occurred with a token is counted and the frequencies of verb frames represent the token. For example:

(1) The fatigue crack growth rate was accelerated by the introduction of one strain cycle.

Parsing sentence 1 shows that for the verb "accelerate", noun phrase "the introduction of one strain cycle" is the first argument and noun phrase "the fatigue crack growth rate" is the second argument. For suffixes of any length of "the fatigue crack growth rate", such as "crack growth rate" or "rate", the second argument of "accelerate", `accelerate-*`, is thus counted.

This clustering scheme is aimed for detecting noun phrases with similar set of co-occurring verb frames. It is proposed under the assumption that co-occurring verb frames are one of the characteristics of technical terms.

Noun phrases and corresponding verb frames are gathered from training set and used in the training phase of the LDA model. A noun phrase in queried document is then matched to noun phrases in training set which shares longest suffix tokens. All tokens in the noun phrase thus include features indicating the cluster. Note that noun phrases which do not match any noun phrases in training set will not include any noun phrase clustering feature. For example, the noun phrase "crack growth rate" will be represented by the sum of frame frequencies of noun phrases "reaction rate", "heat exchange rate", and "birth rate", but if there is no any noun phrase with the suffix "rate", "crack growth rate" will include no noun phrase clustering feature.

4 Experiments

4.1 Dataset

To verify the performance of our method, we performed an experiment on a corpus of 2,079 abstracts retrieved from Scholarly and Academic Information Navigator (CiNii) [1], a scientific paper database provided by National Institute of Informatics (NII). The dataset includes author-assigned keywords which are used to automatically tag the training set.

The corpus is randomly separated into training set of 1,879 abstracts, development set of 40 abstracts, and test set of 160 abstracts. Previous works such as [4, 7] could be evaluated only with precision because the lack of fully annotated corpus. However, our development set and test set are manually annotated by two annotators. The annotation agreement is Cohen's alpha = 0.48. Evaluation with recall score is also vital for term extraction system, since evaluation only with precision score might lead to system capable to recognize only a few trivial terms.

4.2 Models

We test our method by comparing four models:

1. Naive keyword tag (KW).
2. CRF on baseline features (B), which are context features and word-shape features. Context features are word surface and part-of-speech tag up to trigram in 5-token windows centered on the token in question. Word-shape features are features with templates described in Table 1.
3. CRF on baseline features with document cluster (DOC).
4. CRF on baseline features with noun phrase cluster (NP).

The models KW and B are the baselines. We set number of topic for LDA model for document clustering to 18 and that for noun phrase clustering to 20. The numbers are tuned with the development set.

4.3 Evaluation

We use precision, recall, and F1 score as the evaluation measures and adopt two counting schemes.

Span counting In span counting scheme, a span tagged by the system in the text is considered correct when the corresponding span presents in the gold data.

[1] http://ci.nii.ac.jp/

Feature	Type	Input	Value(s)
Text	Text	Computer	Computer
Lower-cased	Text	NLP	nlp
Prefixes: sizes 3 to 5	Text	language	lan, lang, langu
Suffixes: sizes 3 to 5	Text	language	age, uage, guage
Stem [9]	Text	effective	effect
Is a pair of digits	Bool	12	True
Is four digits	Bool	2012	True
Letters and digits	Bool	SK125	True
Digits and hyphens	Bool	7-11	True
Digits and slashes	Bool	24/7	True
Digits and colons	Bool	3,000	True
Digits and dots	Bool	2.718	True
Upper-case and dots	Bool	H.P.	True
Initial upper-case	Bool	John	True
Only upper-case	Bool	ACL	True
Only lower-case	Bool	grep	True
Only digits	Bool	15089	True
Only non-alpha-num	Bool	%&!	True
Contains upper-case	Bool	eXternal	True
Contains lower-case	Bool	BioNLP	True
Contains digits	Bool	Y2K	True
Contains non-alpha-num	Bool	100%	True
Date regular expression	Bool	2012-01-01	True
Pattern	Text	3-26abC	0-00aaA
Collapsed Pattern	Text	3-26abC	0-0aA

Table 1: Word-shape feature templates (adapted from [10])

Type counting In type counting scheme, for each document, the model set is the collection of words tagged by the system, while the gold set is the collection of those annotated manually. The correct instances are words in the intersection of the two sets.

4.4 Tools

Our models are based on CRF and we use CRFsuite [8] [2] implementation of CRF model. As for LDA model, we used the implementation by Blei [3], an author of [2].

We use NLTK [1] for general NLP tasks such as sentence and word tokenization and part-of-speech tagging, while parsing is performed using enju parser version 2.4.2 [4].

5 Result

The result of the experiment with span counting scheme is shown in the Table 2. The NP model gives the highest precision while KW baseline model gives the highest recall and F1 score.

On the other hand, the result of the experiment with type counting scheme is shown in the Table 3. Similar to the previous counting scheme, the NP

[2] http://www.chokkan.org/software/crfsuite/
[3] http://www.cs.princeton.edu/~blei/lda-c/
[4] http://www.nactem.ac.uk/enju/

	P	R	F1
KW	26.81	24.79	25.76
B	26.41	19.11	22.18
DOC	27.03	19.65	22.76
NP	27.06	19.76	22.84

Table 2: Result for span counting

model gives the highest precision while KW baseline model gives the highest recall and F1 score.

	P	R	F1
KW	45.94	18.52	26.39
B	47.61	16.27	24.25
DOC	48.19	17.03	25.17
NP	48.47	17.03	25.21

Table 3: Result for type counting

Since keywords are dictionary information we can use from the test set, it is possible to combine results from the model KW and other models to improve recall while trading off precision. The results of combinations are shown in Table 4 for span counting scheme and in Table 5 for type counting scheme.

	P	R	F1
KW	26.81	24.79	25.76
KW + B	26.72	27.58	27.14
KW + DOC	26.97	28.09	27.52
KW + NP	26.95	27.94	27.43

Table 4: Combined result for span counting

As shown in tables 4 and 5, the precision scores for combined LDA models drop slightly, but are still higher than that of KW model while the overall F1 scores become higher than that of KW model in both counting schemes.

6 Conclusion and discussion

We proposed a simple method to incorporate two different types of LDA-clustering features in a CRF model for term extraction trained on an automatically tagged corpus. We then evaluate our method using manually annotated abstracts of scientific papers. The results show that proposed models could help dictionary method discover more terms with relatively high precision. We thus see possibility of our proposed methods to be used in a bootstrapping system by which the result is expanded in each iteration.

There are many limitations in our works, especially formulation of the problem as sequential labeling. Under current problem formulation, no two overlapping technical terms can be annotated or extracted, although technical terms are often appeared as nested. A revision of the formulation is needed in order to extract all the terms. We also plan to explore the applicability of other extraction methods than CRF to overcome these problems.

	P	R	F1
KW	45.94	18.52	26.39
KW + B	46.18	21.64	29.47
KW + DOC	46.32	22.20	30.01
KW + NP	46.34	22.05	29.88

Table 5: Combined result for type counting

References

[1] Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[2] Blei, D. M., Ng A. Y., and Michael, J. I. Latent dirichlet allocation. *J. Machine Learning Research*, vol. 3, pp. 993-1022. 2003.

[3] Chaimongkol, P., Stenetorp P., and Aizawa, A. Utilising Bilingual Lexical Resources for Technical Term Extraction. *Proc. 26th Annual Conference of the Japanese Society for Artificial Intelligence*, Yamaguchi, Japan, June 2012.

[4] 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.

[5] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th ICML '01*, pp. 282—289, 2001.

[6] Lin, D. and Wu, X. Phrase clustering for discriminative learning. *Proc. Joint Conf. 47th Annual Meeting of the ACL and 4th Int. Joint Conf. Natural Language Processing of the AFNLP*, pp. 1030—1038, 2009.

[7] 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.

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

[9] Porter, M. F. Readings in information retrieval. pp. 313—316. Morgan Kaufmann Publishers, 1997.

[10] Stenetorp, P., Pyysalo, S., and Tsujii, J. SimSem: fast approximate string matching in relation to semantic category disambiguation. *Proc. BioNLP '11 Workshop*, pp. 136—145, 2011.