Method mention extraction from scientific research papers

Hospice Houngbo   Robert E. Mercer
Department of Computer Science
The University of Western Ontario, London, ON, Canada
hhoungbo@uwo.ca, mercer@csd.uwo.ca

ABSTRACT

Scientific publications contain many references to method terminologies used during scientific experiments. New terms are constantly created within the research community, especially in the biomedical domain where thousands of papers are published each week. In this study we report our attempt to automatically extract such method terminologies from scientific research papers, using rule-based and machine learning techniques. We first used some linguistic features to extract fine-grained method sentences from a large biomedical corpus and then applied well established methodologies to extract the method terminologies.

We focus the present study on the extraction of method phrases that contain an explicit mention of method keywords such as (**algorithm, technique, analysis, approach and method**) and other less explicit method terms such as **Multiplex Ligation dependent Probe Amplification**. Our initial results show an average F-score of 91.89 for the rule-based system and 78.26 for the Conditional Random Field-based machine learning system.

KEYWORDS: terminology extraction, rule-based, machine learning, corpus, linguistic features.

Proceedings of COLING 2012: Technical Papers, pages 1211–1222,
COLING 2012, Mumbai, December 2012.

1 Introduction

The methods and techniques used during scientific experiments and reported in scientific papers are often expressed in different forms. They can be in the form of a semantic sequence (see (Hunston, 2008) containing the word *method (“rank-based normalization method” in Sentence 1),* the word *technique (“Non-negative matrix factorization technique” in Sentence 2),* the word *analysis (“discriminant analysis” in Sentence 3),* and so on.

Sentence 1: *In order to compare U133A and U133 Plus 2.0 data we further normalized the data with a **rank-based normalization method**.*

Sentence 2: *In the Bioinformatics field a great deal of interest has been given to **Non-negative matrix factorization technique NMF** due to its capability of providing new insights and relevant information about the complex latent relationships in experimental data sets.*

Sentence 3: *The relative performance of the four IBDQ dimensions in distinguishing best patients with minor symptoms from those with severe was studied by **discriminant analysis**.*

A method mention can be a terminology term such as the phrase **Multiplex Ligation dependent Probe Amplification** in Sentence 4.

Sentence 4: *Recently **< Multiplex Ligation dependent Probe Amplification > < MLPA >** has also been used to quantify copy number classes.*

A method expression can also be a verb phrase referring to an action performed during an experiment, such as the verb phrase *to search for* in sentence 5.

Sentence 5: *These sequences were used **to search for** the nearly invariant nucleotides of the inverse core AAC and core GTT sites separated by a distance typical of previously identified attC sites to bp .*

In some cases, the context in which the method terminologies are used in the text can contain valuable information about the method mention, such as synonyms, definition, and other relations with entities in the text.

The automatic extraction of terminologies has been the focus of many studies in the past (Maynard and Ananiadou, 2000), (Zhang and Wu, 2012), but not many works have focussed on automatic extraction of method mentions. We believe that the extraction of method expressions from research papers can help to build lexical resources that can be used in various NLP tasks on research papers.

For example, the automatic recognition of method expressions can help to easily detect method sentences and classify them into their rhetorical categories as in (Agarwal and Yu, 2009).

In addition, the automatic extraction of method mentions and the information surrounding them can be useful in dictionary building, ontology population and glossary creation. Also, it can help in the task of knowledge discovery from scientific papers, as any mention of methods and techniques used in that paper can easily be presented to readers without them having to read the whole text.

The extraction of contextual information around the method mentions can be useful in building Question-answering and focussed text summarization systems.

In this study, we present our attempt to extract method terminologies from sentences. Also we showed how we can use a terminology context to extract other relevant information about the term. We define “other relevant information” to be the syntactic and semantic relationship between the term and other words. It may be a definition, a variant terminology and so on.

This study doesn’t take into account the extraction of verbs as method mention in scientific research papers.

The contribution of our work is twofold. First we used some linguistic filters to automatically select thousands of fine-gained method sentences that have a high probability to contain methodology terminologies; then we used the context of such sentences to extract terminologies and other information about them.

The rest of this paper is organized as follow. The next section reviews some related works. In Section 3, a detailed methodology of method mentions extraction techniques is presented. In Section 4, the results are described. We conclude the paper by a summary and directions for future work.

2 Related work

Automatic term extraction is an important component of many natural language processing systems. It is often used in applications such as knowledge discovery, knowledge management, automatic text indexing, and so on. Many studies have been conducted on the recognition of terminologies from research papers, especially in the biological domains where new terms are created constantly. These studies primarily focussed on the extraction of domain specific concepts such as nouns and collocations. But it is hasn’t been possible to apply a general rules to extract all the terms.

Previous studies on automatic term extraction have been made possible with the availability of many language resources and large electronic corpora. Three main approaches are used in terminology extraction, namely, linguistics-based approaches, statistical approaches and hybrid approaches.

2.1 Linguistics-based approach

A term from a terminology is often a unit of meaning related to a specific field or domain of study. It can be a single word, such as *clustering*, or a compound expression made up of individual terms, such as *Hidden Markov Model*. Some terms in a terminology can follow comprehensive rules which can help with their generalization.

With the linguistics-based approach, candidate terminology is filtered by linguistic features using morphological analysis, such as part of speech (POS). Complex terms are extracted using shallow parsing and dependency analysis between words in the sentence (Bourigault, 1992). (Dagan and Church, 1994) limited the candidate terminology to a string that represents the pattern of noun sequences. Good results can be achieved in small corpora using linguistic methods, but due to the shortage of patterns, recall can be low and it is difficult to generalize these techniques across fields and languages.

2.2 Statistical approach and machine learning approach

Statistical approaches are based on statistical information, such as the frequency of terms appearing in the corpus. As already noted in (Zhang and Wu, 2012), statistical features can include term document frequency and inverse document frequency TF*IDF (Maedche and Staab, 2000), KF*IDF (Xu et al., 2002), C-value/NC-value (Frantzi et al., 2000) and so on. Term extraction is based on the computation of the Unithood – i.e. a degree of strength or stability of syntagmatic combinations or collocations, and Termhood – i.e. the degree that a linguistic unit is related to a domain-specific concept (Kageura and Umino, 1996) of the terminology.

Termhood calculation is often based on the frequency of the term in the paper or a given baseline corpus.

(Church and Hanks, 1990) used Mutual Information (MI) to compute Unithood, (Dunning, 1993) used LogL , and (Patry and Langlais, 2005) computed left/right entropy to extract unithood candidates. When computing termhood, methods such as TF*IDF (Maedche and Staab, 2000), DR-D (Velardi et al., 2001), C-value/NC value (Frantzi et al., 2000)) and Domain Component Feature Set (DCFS) (Zhang and Sui, 2007) are employed, as reported in (Zhang and Wu, 2012).

2.3 Hybrid approach

Linguistics-based and statistical approaches have their own advantages and disadvantages. They are often integrated to extract terminology. There are two ways to combine them. One way is to extract candidate terms with linguistics methods and then if no terms are found then the statistical methods are applied. (Daille, 1996) used the linguistic methods to get candidate terms and set them as the input of statistical models. Then statistical methods such as MI and LogL are used to get final terms.

The other way is to obtain candidate terms using statistical methods first and then use linguistic methods to discard those terms that are inconsistent with linguistic patterns. (Maynard and Ananiadou, 2000) extracted multi-word terms using a thesaurus and the semantic Web concept to get the semantic and category information, and then integrated it with the statistical and syntactic information in the corpora. Different terminology extraction toolkits can also be integrated to extract terminology besides the integration of linguistics and statistics methods. (Vivaldi and Rodriguez, 2001) integrated different term extraction tools using simple voting and the results were better than that with single term extraction tools. (Vivaldi et al., 2001) improved the previously mentioned voting approach, got the best integration strategy with Boosting algorithm and improved the performance of terminology extraction based on the hybrid approach.

3 Methodology

3.1 Corpus creation

The goal of this study is to extract the methods and techniques used in biomedical research papers in order to build a lexical resource comprising the name, the variants and definition of such methods and techniques as they are mentioned and used in research articles. The first task consists of the gathering of a comprehensive corpus that is large enough to contain an important number of method mentions and information about how they are used in the papers. One way is to select some research papers and scan through each of them to extract candidate "method sentences" manually. Another way is to use filters to automatically select in a large repository of papers, "method sentences" that we believe are about the methods or techniques used in the paper as well as the context of such sentences. We define the context of a "method sentence" to be a window of sentences coming after or before them. The first option has proved in the past to be more precise, but very difficult to implement. In fact, a manual information extraction task can be tedious and time-consuming and requires many specialized skills and domain experts. It is therefore recommended to use automatic extraction techniques whenever it is possible. Since our purpose is to find as many relevant sentences as possible, it is obvious that we cannot rely on existing corpora to achieve our goal.

Some recent works have focussed on the classification of sentences from biomedical articles into the IMRAD (Introduction, Methods, Research, and, Discussion) categories. In their attempt to classify biomedical research papers into these categories (Agarwal and Yu, 2009) used a corpus of 1131 sentences. Of these sentences, 389 were labelled Introduction, 363 were labelled Methods, 273 were labelled Results and 106 were labelled Discussion. Even though the corpus contained "method" sentences, very few contain the mention of the techniques used in the paper. Also, the context of the sentence was not retrieved. (Liakata et al., 2012) used a corpus of 265 articles from biochemistry and chemistry annotated at the sentence level by experts in the domains. Even though the corpus contains 8404 method sentences, many sentences belong to the same papers and are about the same methods or techniques. For instance 10 consecutive sentences can be annotated to belong to the method category. Those sentences usually refer to the same method or technology and some of them may not even contain any mention of a method. Besides, few of the sentences contain a definition of the techniques used in the papers.

Based on the study of the corpora mentioned above, we deemed it necessary to build a different corpus that contains as many method sentences as possible and a definition or a usage context of the method mention.

We relied on some linguistic concepts such as anaphoric relations to produce our fine-grained method corpus.

In fact, most demonstrative noun phrases are anaphoric; therefore a sentence that begins with the demonstrative noun phase "This method" is anaphoric and its antecedents are likely to be found in previous sentences. (Torii and Vijay-Shanker, 2005) reported their work on anaphora resolution of demonstrative noun phrases in Medline abstracts, and found that nearly all antecedents of such demonstrative phrases can be found within two sentences. On the other hand, (Hunston, 2008) reported that interpreting recurring phrases in a large corpus enables us to capture the consistency in meaning as well as the role of specific words in such phrases. So, the recurring semantic sequence "this method" in the Pubmed corpus can help us to capture valuable information in the context of their usage.

To build our corpus we therefore search for sentences starting with "This method" in the PubMed article repository as well as the sentences immediately preceding them. Then we collected the pairs of sentences.

Sentence 6 in Example 1 contains the mention of the method and Sentence 7 contains its usage and definition context.

Example 1

Sentence 6 *: The Ortholuge method reported here appears to significantly improve the specificity*

(precision) of high-throughput ortholog prediction for both bacterial and eukaryotic species .

Sentence 7 *: This method , and its associated software , will aid those performing various comparative genomics-based analyses , such as the prediction of conserved regulatory elements upstream of orthologous genes .*

We can see that Sentence 6 is talking about the method and Sentence 7 is a reference to the method mention and how it is used. Combining both sentences we therefore have sufficient information to extract the “method” mention, its usage and its benefits.

We can then derive such information to fill lexical components such as:

- Method Mention: ***Ortholuge method***
- Usage/Role: ***comparative genomics-based analyses/ prediction of conserved regulatory elements upstream of orthologous genes.***

Example2

Sentence 8 *: An alternative method for predicting protein function is the Phylogenetic profile method, also known as the Co-Conservation method , which rests on the premise that functionally related proteins are gained or lost together over the course of evolution [4] .*

Sentence 9 *: This method predicts functional interactions between pairs of proteins in a target organism by determining whether both proteins are consistently present or absent across a set of reference genomes.*

In Example 2, it is possible to extract the following information:

- Method : ***Phylogenetic profile method***
- Variant/ also known as:***Co-Conservation method***
- Usage: ***for predicting protein function / predicts functional interactions between pairs of proteins***
- How: ***by determining whether both proteins are consistently present or absent across a set of reference genomes.***

The information that we can derive from this pair is sufficient enough to create lexical resources that can be used in many natural language processing tasks.

Using the retrieval technique mentioned above, we have been able to retrieve about 6500 such pairs of sentences from 189 different journals and 2000 papers.

We limit the scope of this study to the extraction and recognition of the method terminologies, due to time constraints.

3.2 Gold Standard datasets

We created 2 sets of gold standards with sentences taken only from BiomedCentral journals. The first gold standard comprises 918 pairs of sentences containing the first category of method mention – i.e. terminology units ending with a method keyword. The second gold standard comprises 122 pairs of sentences of method mentions that don’t contain a method keyword. In each gold standard, we assumed that the method mention is in the first sentence and the other information about its usage is in the second sentence.

We used grammatical rules to extract the first category (method mentions that contain keywords, such as algorithm, technique, analysis, approach and method) and machine learning techniques to extract the second category of method mention (those that don't contain the above keywords).

When a method keyword is not explicitly mentioned in a sentence containing a method mention, it is not obvious to apply general grammatical rule to recognize it as the words composing it can be of various forms. In the following sentences:

Sentence 10 : Enault and colleagues proposed an improved **< phylogenetic profile >** based on a **< normalized Blastp bit score >**.

Sentence 11 : Another way to obtain suboptimal solutions from a **< HMM >** is to do **< HMM sampling >**.

Sentence 12 : In this paper we introduce a **< Bootstrap procedure >** to test the null hypothesis that each gene has the same relevance between two conditions where the relevance is represented by the Shapley value of a particular coalitional game defined on a microarray data set .

we can notice that the different method mentions - *< phylogenetic profile >, < normalized Blastp bit score > < HMM >, < Bootstrap procedure >* are all of different word shapes. Also when most of them are nouns phrases, they can be confused with other noun phrases in the sentence. We thus believed that the extraction of this type of method mentions can be viewed as a named entity recognition task (Maynard et al., 2001), (Palmer and Day, 1997).

For this second task, we transformed each sentence into the BIO format (Table 1). There are 284 manual-tagged terms, 122 sentences, 2871 words and punctuations.

In Table 2 we shows the number of sentences in each dataset.

Enault and colleagues proposed an improved **< phylogenetic profile >** based on a **< normalized Blastp bit score >**	Enault	O
	and	O
	colleagues	O
	proposed	O
	phylogenetic	B-method
	profile	I-method
	based	O
	on	O
	a	O
	normalized	B-method
	Blastp	I-method
	bit	I-method
	score	I-method
	.	O

Table 1: Representation of a method sentence in the BIO format.

Category (keywords)	Number of sentences	Proportion
Method	439	42%
Analysis	200	19%
Model	63	6%
Algorithm	73	7%
Approach	145	14%
Other (Machine learning corpus)	122	12%
Total	1040	100%

Table 2: Corpus statistics (combining both datasets).

3.3 Rule-based extraction

Most of the method mentions in the first category can be represented by the following examples:

1. *rank-based normalization method*
2. *HRV power spectral analysis*
3. *Non-negative matrix factorization technique*
4. *linear regression analysis*
5. *Newton-type algorithm*
6. *tube group amplification approach*
7. *progressive alignment algorithm*
8. *metabolite profiling approach coupling mass spectrometry*
9. *profile-based HMM method*
10. *multifactor-dimensionality reduction MDR method*

As we can notice these are simple grammatical patterns that can be extracted with simple rules.

1. They can start and continue either with an adjective or a noun.
2. They can continue either with an adjective or a noun.
3. They end with a method keyword.

These rules can be represented by the following regular expression:

(Adjective | Noun)+(method | analysis | algorithm | approach| model)

To extract such patterns, we first used the Genia tagger (Tsuruoka et al., 2005), a Part-of-Speech tagger trained on a biomedical corpus, to tag every word in the sentence. Then, we used the rules to extract all phrases and terminologies that correspond to the above mentioned patterns. The results are presented in the Result section.

3.4 Machine learning and feature extraction

3.4.1 Feature extraction

The feature pool includes:

1. Word feature

   The word itself.

2. Part-of-speech tags

   We used a POS tagger (the Genia tagger) to tag each word in the sentence.

3. Word-shape features

   We check whether the word is lower case, upper case or has both lower case and upper case letters.

   These features include: isAllCaps, StartWithCap, isAllLowerCase, isMixedCase.

4. Position features

   We checked if the word is at the beginning of sentence (BOS), at the end of sentence (EOS), Not Beginning of Sentence (!BOS), Not Ending of Sentence (!EOS).

5. Token prefixes and suffixes features

   We extract the prefixes and suffixes for each word. These include the first four prefixes and the last four prefixes for each word.

6. Bigram features

   For each sentence we extracted bigrams containing only nouns and adjectives.

3.4.2 Conditional Random Field (CRF)

To train the model, we used Conditional Random Field machine learning on 90 % of the dataset and we tested on 10 % of the dataset.

CRF is a machine learning model proposed by (Lafferty et al., 2001). It is widely used in word segmentation, part-of-speech tagging, chunking recognition, named entity recognition and so on.

Conditional Random Fields (CRF) model is a state-of-the-art sequence labeling method, which can use the features of documents more sufficiently and effectively. As we said earlier, non-explicit method mention extraction can be considered as a task of entity recognition to which string labeling techniques can be successfully applied.

4 Results and Discussion

The experimental results used 924 for the rule-based task and 122 sentences for the machine learning task. Table 3 shows the performance of the two tasks. As we could expect recall is very high (100 %) for the rule-based task because every sentence in the dataset contains a method keyword. Our rules were able to recognize every noun phrase and adjective phase ending with a method keyword. Precision is 85.40, which is not bad. Some of the errors came from the tagger. We have spotted these errors and we will remove them in future study For the machine learning task, precision is 81.8 percent, recall is 75 and F-score is 78.26 percent. Features such as word shape, POS, and noun-bigrams performed better than the position features. Also some of the errors came from the all-lower-case terms as they tend to be confused with similar words in the sentence. Table 3 shows the performance of both systems. We believe that the scores can be improved with better linguistic filters in the case of the first task and a better feature selection for the second task.

System	Precision	Recall	F-Measure
Rule-based	85.40	100	91.89
Machine Learning	81.8	75.00	78.26

Table 3: Precision, Recall, F-measure of the Various Methods.

Our work can be compared to (Zhang et al., 2008) which uses CRF for automatic keyword extraction from documents; they reported promising results (F-score of 51.25 percent) on a Chinese corpus.

It can also be compared to (Zhang and Wu, 2012), which uses a multi-level termhood method to extract terminology candidates from a bilingual corpus. Their system achieves an F-score of 79.6% with CRF. Both works use similar techniques to ours, but most of the terminology extraction tasks are performed using a Chinese corpus.

5 Conclusions and future work

In this paper we have first presented a simple approach to extract fine-grained method sentences from large scientific corpora. We have also explored two established techniques to automatically extract method terminologies from method sentences. Our results showed that we can extract most of these terms using simple grammatical patterns. A few other terms can be extracted with machine learning techniques. A brief study of the corpus showed that the context of the method mentions can help in the extraction of important information about the method term. Our future work will then be to use the whole corpus to extract such information that is essential in the building of NLP resources such as glossaries, ontologies and specialist lexicons.

References

Agarwal, S. and Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics - Volume 3*, COLING '92, pages 977–981, Stroudsburg, PA, USA. Association for Computational Linguistics.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Dagan, I. and Church, K. W. (1994). Termight: Identifying and translating technical terminology. In *ANLP*, pages 34–40.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Klavans, J. and Resnik, P., editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 29–36. MIT Press, Cambridge, MA.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Hunston, S. (2008). Starting with the small words patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics*, 13(3):271–295.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C. R., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maedche, A. and Staab, S. (2000). Mining ontologies from text. In *Proc. of Knowledge Engineering and Knowledge Management (EKAW 2000)*, LNAI 1937, pages 169–189. Springer.

Maynard, D. and Ananiadou, S. (2000). TRUCKS: a model for automatic multi-word term recognition. volume 8, pages 101—125.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. (2001). Named entity recognition from diverse text types. In *In Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark*.

Palmer, D. D. and Day, D. S. (1997). A statistical profile of the named entity task. In *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97, pages 190–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patry, A. and Langlais, P. (2005). Corpus-based terminology extraction. In *7th International Conference on Terminology and Knowledge Engineering*, pages 313–321, Copenhagen, Denmark.

Torii, M. and Vijay-Shanker, K. (2005). Anaphora resolution of demonstrative noun phrases in medline abstracts. In *Procedings of PACLING 2005*, pages 332–339.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and ichi Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In Bozanis, P. and Houstis, E. N., editors, *Panhellenic Conference on Informatics*, volume 3746 of *Lecture Notes in Computer Science*, pages 382–392. Springer.

Velardi, P., Missikoff, M., and Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, pages 5:1–5:8, Toulouse.

Vivaldi, J., Màrquez, L., and Rodríguez, H. (2001). Improving term extraction by system combination using boosting. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 515–526, London, UK. Springer-Verlag.

Vivaldi, J. and Rodriguez, H. (2001). Improving term extraction by combining different techniques. *Terminology*, 7(1):31–48.

Xu, F., Kurz, D., Piskorski, J., and Schmeier, S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In *3rd International Conference on Language Resources and Evaluation*, page 7pp. European Language Resources Association.

Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., and Wang, B. (2008). Automatic keyword extraction from documents using conditional random fields. In *Journal of Computational Information Systems*, pages 1169–1180. Binary Information Press.

Zhang, C. and Wu, D. (2012). Bilingual terminology extraction using multi-level termhood. *The Electronic Library*, 30(2):295–309.

Zhang, Q. L., Q. L. and Sui, Z. F. (2007). Measuring termhood in automatic terminology extraction. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pages 328–335. IEEE Press, Piscataway, NJ.