Qual Quant (2017) 51:1009–1025
DOI 10.1007/s11135-016-0369-3

A semantic annotation framework for scientific publications

Yuchul Jung[1]

Published online: 11 June 2016

Abstract Considering the growing volume of scientific literature, techniques that enable automatic detection of informational entities existing in scientific research articles may contribute to the extension of scientific knowledge and practical usages. Although there have been several efforts to extract informative entities from patent and biomedical research articles, there are few attempts in other scientific literatures. In this paper, we introduce an automatic semantic annotation framework for research articles based on entity recognition techniques. Our approach includes tag set modeling for semantic annotation, semi-automatic annotation tool, manual annotation for training data preparation, and supervised machine learning to develop entity type recognition module. For experiments, we choose two different domains, such as information and communication technology and chemical engineering due to their high usages. In addition, we provide three application scenarios of how our annotation framework can be used and extended further. It is to guide potential researchers who are willing to link their own contents with external data.

Keywords Entity type recognition · Research article · Structural support vector machine · Semantic annotation · Knowledge construction

1 Introduction

As responses to global open data movement, various efforts have been explored to collect and share public data. However, open data is mostly focused on non-textual material such as genomes, chemical compounds, maps, medical data, etc. In aspects of understanding social issues or accidents (e.g., earthquake and sunken ship), currently available public open data are hard to be linked to scientific knowledge or factual evidences that are helpful

✉ Yuchul Jung
enthusia77@gmail.com; jyc77@kisti.re.kr

[1] Department of Information Convergence Research, Korea Institute of Science and Technology Information (KISTI), 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea

for understanding the accidents because they were made with their own purposes without the consideration of collaboration with other domains' knowledge. Therefore, we are aware of few use cases in science areas due to the lack of trustable, linkable scientific knowledge.

From a business and government point of view, there is an increasing need to use large-volume media, such as Facebook, Twitter, and Web news because they contain various types of information and responses surrounding our human lives. However, knowledge obtained from online web sites and social media comes with a major caveat—it cannot always be trusted, nor is always factual evidence of high quality.

Meanwhile, excavating technology opportunities from patent data is one of promising usages which have shown the benefits of currently available large-scale, trustable data (Yoon et al. 2014). In addition, (IBM Watson 2014) show the successful cases of combining clinical trial outcomes in pharmaceuticals and automatically extracted knowledge from biomedical research articles. By constructing an information map based on the reasoning results from the clinical outcomes and experimental knowledge obtained from numerous research articles, it can infer possible side effects of a drug.

At any rate, to expedite knowledge construction and sharing between industry-government-academic (i.e., triple helix), building a trustable knowledge ecosystem where a number of active experts are participating to produce trustable resources is crucial. In addition, considering that technological change may be connected to new products and services, the triple-helix of inter-sectoral cooperation is also important (Phillips 2013).

In this line of context, assuming that scientific research articles are authored by various domain experts and scientists, we are interested in how to build a knowledge construction framework that can be applied to generate linkable, reusable knowledge. Due to the immature stage of user-generated contents on social media, we choose scientific research articles as our main source of knowledge construction.

In this paper, as a way to construct linkable knowledge, we introduce the automatic annotation framework designed for extracting knowledge from various types of scientific research articles. Starting from tag set design for scientific research articles, we fulfil the necessary procedures, such as semi-automatic annotation tool and machine learning based entity classifier. They are to alleviate the difficulties of cost-inefficient data annotation and provide a robust entity identification basis without too much concern about machine learning based training.

Our main goal in this paper is to help other domains' researchers who want to build their own knowledge base and link with external open data. We report our experiences on chemical engineering and information & communication technology domains. In addition, for more extensive understanding, we introduce three promising application scenarios which utilize our constructed knowledge or can be realized by extending our annotation framework in aspects of external linkage and extensibility. We have a firm belief that it can be extended further for different science domains and other user-generated contents only if they maintain reliability.

2 Background

2.1 Entity recognition and disambiguation

Recently, the ERD 2014: entity recognition and disambiguation challenge (Carmel et al. 2014) was held to advance the state of the arte NLP techniques for both short documents

and long documents. The goal of entity recognition and disambiguation (ERD) is to identify mentions of entities and link them a relevant entry in and external knowledge base, such as Wikipedia, DBpedia, Freebase, and so on. The ERD process consists of two steps—spotting and disambiguation—in general. First, the system detects potential mentions of entities in text and links them to a list of senses which are the candidate entities that can be referred to by the given mention. Second, the system disambiguates the candidate entities by selecting the most appropriate entity for each mention. Depending on implementations and purposes, there have their own characteristics (Ferragina and Scaiella 2010; Chiu et al. 2014; Daiber et al. 2012), but they used publicly available large-scale knowledge based as their resources.

In Korean language, a fine-grained named entity recognition (NER) using conditional random fields (CRFs) for question answering (QA) (Lee et al. 2006) was studied to classify 147 classes. It used CRFs to detect boundary of named entities and maximum entropy (ME) to classify named entity classes for time efficiency in training time. However, the wide range of named entity tag set doesn't seem to easily applicable to different kinds of texts due to high dependency to QA domain. As an open domain NER, (Park et al. 2014) proposed title NER using Wikipedia and abbreviation generation. To cover various types of entertainment content on the Internet, the study constructed ten classes, such as title, song, program, movie, drama, cartoon, novel, animation, book, and essay.

Meanwhile, the excessively long training has been a bottleneck in NERs modeling. To resolve it, (Lee et al. 2011) presented a NER using a modified Pegasos algorithm for structural SVMs. Their experiments on 15 classes (e.g., person, location, organization, artifacts, etc.) show that the training time for the modified Pegasos algorithm is decreased about 20 times compared to CRFs.

2.2 Semantic processing for scientific research articles

As an effort to categorize the detailed regions of papers' abstracts, argumentative zoning (AZ) scheme is suggested (Teufel and Moens 2002). AZ is a scheme which provides an analysis of the rhetorical progression of the scientific arguments, following the knowledge claims, such as AIM, TEXTUAL, BASIS, CONTRAST, BACKGROUND, OTHER, and OWN. The scheme was applied to computations linguistics papers and revised for biology papers (Mizuta et al. 2006), chemistry and computational linguistics papers (Teufel and Batchelor 2009). For example, the categories of AZ in biomedical abstract can be BACKGROUND, OBJECTIVE, METHOD, RESULT, CONCLUSION, RELATED_WORK, and FUTURE_WORK. To enable automatic AZ, positional heuristics, semantic patterns, and weakly supervised variations were developed (Ibekwe-SanJuan 2010; Guo et al. 2011). The results of AZ can be utilized for question & answering, summarization, and semantic publishing.

(Ibekwe-SanJuan 2010) tested two different approaches—linguistic cues and positional heuristics—to develop automatic annotation techniques for argumentative roles of sentence sin scientific abstracts. The argumentative roles are OBJECTIVE, NEWTHING, RELATED_WORK, RESULT, HYPOTHESIS, and FUTURE_WORK. Through experiments on medline abstracts, it is turned out that positional heuristics perform better than linguistic cues because sentences from different argumentative roles are not always discriminated by surface linguistic cues.

Slightly different from AZ approaches, (Gupta and Manning 2011) proposed a framework to extract three kinds of detailed information, such as focus (main contribution), technique (method or tool used), and domain (application domain) for scientific articles.

They extract the three information by matching semantic patterns to dependency trees and learn the patterns using bootstrapping. They aim to use this rich information obtained from scientific articles to study the dynamics of research communities and to define a new way of measuring influence of one research community on another. They introduced a case study on the computational linguistic community which delivers an observation that speech recognition and probability theory have the most seminal influence. In case of weakly-supervised approach to AZ (Guo et al. 2011), active learning and self-training were combined by making a goody use of both labeled and unlabeled data.

2.3 Semantic annotation and effective usages

More recently, (Tateisi et al. 2014) introduced a new annotation scheme for formalizing typical schemas for representing relations among concepts in research papers, such as techniques, resources, and effects. Their research goal is to build a framework for representing the semantics of research papers in implementing intelligent search systems. Their defined tag set consists of 16 relation types (e.g., APPLY_TO, SUBCONCEPT, RESULT, CONDITION, EVALUATE, etc.) and three entity types (i.e., OBJECT, MEASURE, TERM). The characteristics of the schema are universal in contemporary science and engineering, and sociology. Their experimental results on computer science abstracts written in Japanese showed fairly good agreement in entity and relation type determination. However, machine learning based entity and relation type recognition is remained as their future work. We refer their annotation scheme to derive our entity types.

As a noticeable example of effective use of those entities and relations extracted from scientific researches, we can refer to IBM's Watson Discovery Advisor (2014). It constructs a map of information by reasoning over patterns it "sees" in available data and knowledge. To make the blur metaphor into trustable knowledge, it connects the dots between pieces of related information. Its application areas cover law, pharmaceuticals, biotech, education, chemical, engineering, scientific research, etc. For example, Watson reads and understands research articles that discuss clinical trial outcomes in pharmaceuticals. Based on the new knowledge, IBM Watson can help researchers match a drug with a patient effectively by avoiding side effects. Moreover, it may beneficial to not only professionals but also less experience people by guiding them learn best practices, answering their questions, and helping them to discover patterns from target area's data and knowledge.

2.4 Semantic annotation framework

Considering diverse content areas in the science domain, the cornerstones for the robust entity extraction platform for scientific publications are the construction of trustworthy annotation data that enable high performing entity recognition models and the simplification of the training/testing tasks. Our efforts are focusing on alleviating the following difficulties: (1) domain-specific manually annotations tend to be cost-inefficient and ambiguous even if domain experts involve and (2) training & testing based on the constructed annotation data have been a critical issue for advanced entity recognition system.

2.5 Overall process

As shown in Fig. 1, our suggested semantic annotation framework consists of five steps—tag set design, semi-automatic annotation tool, annotation data construction, structural

SVMs based entity recognition, and experiments. In our approach, tag set design and manual data construction steps require expert-level experience and background knowledge. Only step 4 and 5 can be done automatically based on predefined settings. Meanwhile, annotation tool is built by using the existing open source software (Stenetorp et al.2012) and domain-specific dictionaries as in Table 2.

2.6 Tag set design

The current tag set has five entity types. An entity is whatever can be an argument or participant in a relation which signifies research domain events. Entity types are MEAURE, TERM, TOPIC, METHOD, and EXPERIMENT as shown in Table 1. MEASURE, TERM, and TOPIC are mostly used to annotate mostly nouns or noun phrases. Meanwhile, METHOD and EXPERIMENT are related to 'verb' containing expressions, such as "designed ∼", "performed ∼ experiments", and "utilized such methods ∼". It is reasonable to assume that those entity types have corresponding or linkable relation types, but we do not consider relation modeling in this paper because we only focus on the automatic entity recognition.

The five entity types were extended by revising the existing approaches (Teufel and Batchelor 2009; Tateisi et al. 2014; Eltyeb and Salim 2014; Atdağ and Labatut 2013; Murphy et al. 2006 to cover our target scientific content areas, such as information technology and chemical engineering. At initial design, we included TARGET (subject of application) and SUBCONCEPT (underlying concepts which is highly related with method) as one of our entity types. However, they were removed now due to their high ambiguity. In many cases, their occurring position and syntactic patterns in sentences are not so distinguishable.

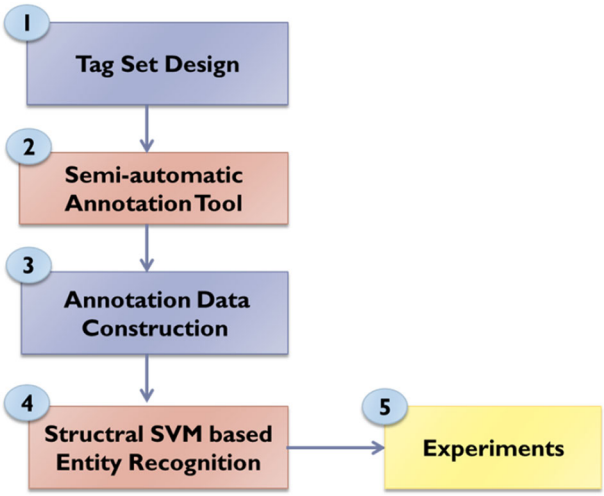

Fig. 1 Overall process

Table 1 Definition of entity types

Entity Type	Definition
MEASURE	Number, numeric measures, parameters used in experiments
TERM	Technical terms and domain keywords that are not selected as topic
KEYWORD	Keywords that are related with the title and main topics of the given paper
METHOD	Means (or methods) of the topic keywords adopt
EXPERIMENT	Experiment results or methods used for the experiment

2.7 Semi-automatic annotation tool

To help annotation task of annotators, we devised a semi-automatic annotation tool based on the ***brat*** annotation tool (Stenetorp et al. 2012). Different from the original ***brat*** tool, our annotation tool support automatic pre-processing based on two domain dictionaries. Figure 2 shows the annotation results for the abstract written in Korean. With the start of the ***brat*** tool, it loads two domain dictionaries into memory. When the target abstract is chosen, it performs keyword spotting task across multiple words to see if the compared words are contained in the dictionaries. More than 20 candidate noun terms were automatically identified and highlighted for further revision by annotators.

For our prototype system, we construct bilingual domain dictionaries for information & communication technology (ICT) and chemical engineering by referring to existing IT web dictionaries (http://word.tta.or.kr/) and language resources provided by Korean Chemical Society (http://new.kcsnet.or.kr/iupacname). Each domain has about 3000 index terms. More specifically, we categorized the dictionaries into four types as in Table 2. The total size amounts to 7213 terms. The dictionaries should be extended further because the

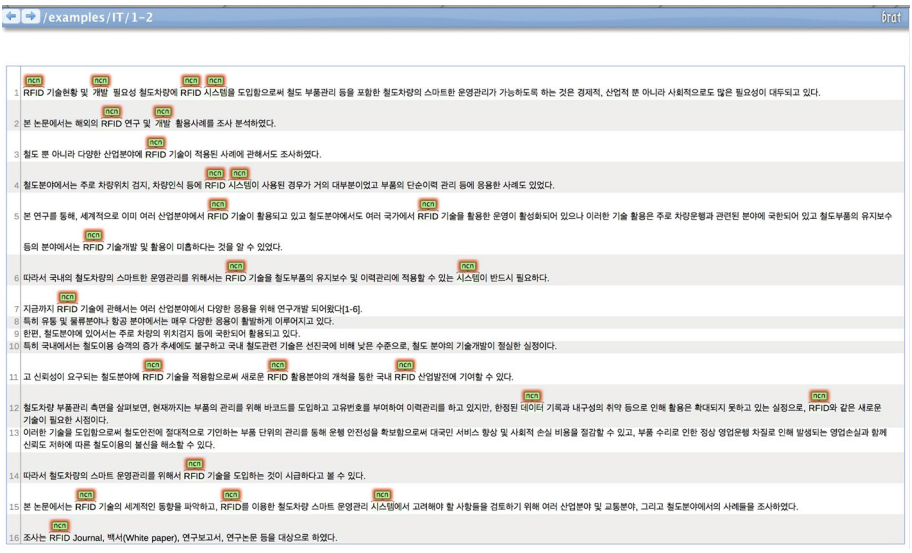

Fig. 2 Dictionary-based automatic noun identification in Korean

numbers are not quite enough to cover all terms appearing in research articles of ICT and chemical engineering. The automatic identification highly depends on the size of dictionaries. In order to adapt different scientific areas, dictionary construction for the target domain is prerequisite.

After selecting certain words or phrases, annotators can choose one type among predefined entity types as in Fig. 3. Revising already tagged expression is also possible. In addition, the annotators can annotate a relation between two different entities. Basically, the ***brat*** tool supports entity type annotation and relation type annotation through the UI based annotation environment.

2.8 Annotation data construction

For the construction of high quality annotation data, which will be used for training and testing the NER models, we utilize research articles' abstracts that have author-defined topic keywords of both in English and in Korean. We then randomly select 1000 papers from the top-ranked journals of ICT and chemical engineering. This is to overcome the difficulties of constructing annotation data which require huge manual efforts in annotation and validation.

Author defined keywords provide quite useful knowledge which is helpful for understanding the given research article. Moreover, they deliver substantial clues for discovering related knowledge in the paper. In this line of context, different from other approaches for research article annotations (Tateisi et al. 2014; Murphy et al. 2006), we only consider KEYWORD entity to reveal the hidden lexico-syntactic patterns of topic related keywords that usually characterize main topics of papers.

2.9 Structural SVM based entity recognition

Learning an entity recognition model is to teach how to annotate the previously defined entity types for an input document. Early studies mostly used handcrafted rules, supervised machine learning approaches which use hidden markov models (Bikel et al. 1998), maximum entropy models (Grishman et al. 1998), and support vector machines (Asahara and Matsumoto 2003). Recently conditional random fields (McCallum and Li 2003) were widely chosen to solve various types of entity recognition systems based on training data according to domains.

Although CRFs is a good solution for building an entity recognition system, we select a structural SVM (Joachims et al. 2009) which outperforms the CRFS in both speed and precision. Structural SVM is a large-margin method for structured output prediction.

Table 2 Domain dictionaries

Domain	Dictionary type	# of terms
Chemical engineering	Chemical formula	1144
	Chemical elements and compound	1044
	Chemical academic jargon	1672
ICT	Information & communication technology jargon	3353
Total		7213

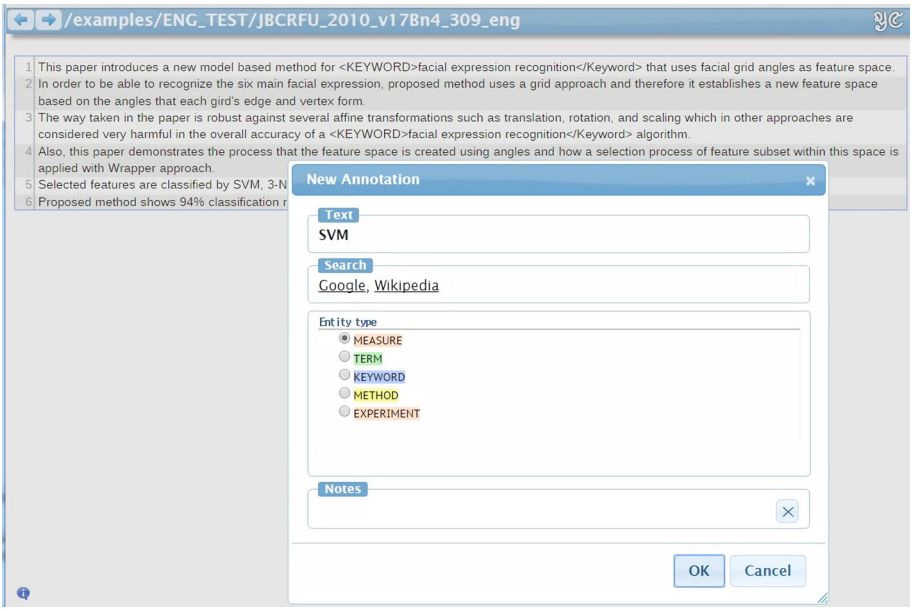

Fig. 3 Manual entity annotation in English using brat tool

1-slack formulation of structural SVM algorithm is faster than existing methods such as SVM-light and sequential minimal optimization (SMO). Its convergence rate is $Q(1/\varepsilon)$.

Structured classification is a task of predicting y from x in cases where y has a meaningful internal structure. For example, x might be a word string and y is a sequence of part of speech labels. Alternatively, y might be a parse tree of x. The approach is to learn the discrimination function f: $X \times Y \rightarrow R$ over $<$ input, output $>$ pairs from which we can derive a prediction by maximizing f over the response variable for a specific given input x. In our structural SVM based learning, we assume f to be linear in certain combined feature representations of inputs and outputs $\Psi(x, y)$, $f(x, y; w) = w^{\mathrm{T}}\Psi(x, y)$. The specific form of $\Psi(x, y)$ depends on the nature of the problem. An example of entity recognition is shown below.

In our entity recognition, we assume that $x = <x_1, x_2, \ldots, x_{\mathrm{T}} >$ be a sequence of words in sentences, and $y = <y_1, y_2, \ldots, y_{\mathrm{T}}, >$ be sequences of labels, each of them are tags associated an entity type and a boundary tag (i.e., B-KEYWORD, I-KEYWORD, B-METHOD, I-METHOD, O). $\Psi(x, y)$ is different according to problem's characteristics.

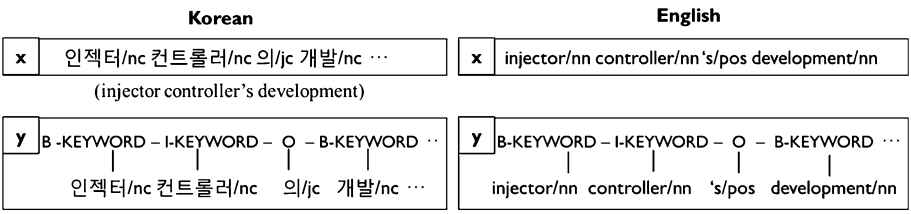

Fig. 4 An example of entity recognition

Figure 4 shows examples of Ψ (x, y) for entity recognition in Korean and in English, respectively. Because part-of-speech taggers follow their own tag sets according to language, the tagged results are somewhat differ each other.

To deal with structured prediction problems in which |**y**| is very large, slack rescaling and margin rescaling were proposed (Tsochantaridis et al. 2004). More specifically, Joachims et al. proposed 1-slack formulation to structural SVMs (Joachims et al. 2009). Its key idea is to replace the n cutting-plane models of each hinge loss with a single cutting plane model of each hinge loss with a single cutting plane model for the sum of the hinge losses. Since there is only a single slack variable, it is referred as "1-slack" structural SVMs. Joachims et al. showed that the dual form of the 1-slack formulation has a solution that is extremely sparse with the number of non-zero dual variables independent of the number of training examples. The margin rescaling formula of 1-slack structural SVMs for entity recognition is as follows:

$$\min_{w,\xi} \frac{1}{2}\left\|w - w_0^2\right\| + C\xi, \mathrm{st}\xi \geq 0$$

$$\forall(\hat{y}_1, \ldots, \hat{y}_n) \in Y^n : \frac{1}{n} W^T \sum_{i=1}^{n} \delta\Psi_i(x_i, \hat{y}_i) \geq \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) - \xi$$

where C is a regularization parameter, $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is a training example, $\delta\Psi_i(\boldsymbol{x}_i, \boldsymbol{y}) = \Psi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Psi(\boldsymbol{x}_i, \boldsymbol{y})$, $\Psi(\mathbf{x}, \mathbf{y})$ is a feature vector function, and L(**y**, **y** is a hamming loss function that counts the number of mislabelings for output **y** relative to the correct output, $\boldsymbol{y}_i$. Unlike regular SVMs, structural SVM can be used for predicting complex y outputs such as trees, sequences, or sets. For training 1-slack structural SVM, 1-slack cutting-plain algorithm can be basically used (Joachims et al. 2009). The pseudo code of the algorithm is shown in Fig. 5.

In our implementation, we chose to use the modified fixed-threshold sequential minimal optimization (FSMO) for 1-slack structural SVM problems (Lee and Jang 2010). The modified FSMO is easy to implement and conceptually simple because it assumes that the formulation of 1-slack structural SVMs has no bias and no linear equality constraint binary classification of SVMs. If data set is quite large and high speed is the principal

Fig. 5 1-slack cutting-plane algorithm (excerpted from Joachims et al. 2009)

Algorithm 1. 1-Slack Cutting Plane Algorithm

1: Input: $(\mathbf{x}_1,\mathbf{y}_1), \ldots, (\mathbf{x}_n,\mathbf{y}_n), C, e$
2: $S \leftarrow \emptyset$
3: **repeat**
4: $(\mathbf{w},\xi) \leftarrow \arg\min_{\mathbf{w},\xi>0} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi$
$s.t. \forall(\hat{\mathbf{y}}_1,\cdots,\hat{\mathbf{y}}_n) \in S:$
$\frac{1}{n}\mathbf{w}^T\sum_{i=1}^{n}\delta\Psi_i(\mathbf{x}_i,\hat{\mathbf{y}}_i) \geq \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i,\hat{\mathbf{y}}_i) - \xi$
5: **for** i=1,...,n **do**
6: $\hat{\mathbf{y}}_i \leftarrow \arg\max_{\mathbf{y}\in Y}\{L(\mathbf{y}_i,\mathbf{y}) + \mathbf{w}^T\Psi(\mathbf{x}_i,\mathbf{y})\}$
7: **end for**
8: $S \leftarrow S \cup \{(\hat{\mathbf{y}}_1,\cdots,\hat{\mathbf{y}}_n)\}$
9: **until** $\frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i,\hat{\mathbf{y}}_i) - \frac{1}{n}\mathbf{w}^T\sum_{i=1}^{n}\delta\Psi_i(\mathbf{x}_i,\hat{\mathbf{y}}_i) \leq \xi + e$
10: return $(\mathbf{w}, \xi)$

requirement, the extended Pegasos algorithm for structural SVMs (Lee et al. 2011) is recommendable. In the algorithm, the objective of Pegasos is replaced with an approximate objective function by setting $\lambda = 1/C$ (C is a regularization parameter).

For our proposed entity type recognition, the following features were used.

(1) Word feature: orthogonal features for words at position $\{-2, -1, 0, 1, 2\}$
(2) Suffix feature: suffixes of length $1 \sim 3$ for words at position $\{-2, -1, 0, 1, 2\}$
(3) Part-of-speech (POS) tag feature: POS tag for words at position $\{-2, -1, 0, 1, 2\}$
(4) List lookup feature: list of entities and list of entity cues
(5) Character level regular expression feature: such as [A–Z]*, [0–9]*, [0–9], [0–9], [0–9], [0–9], [0–9], [A–Za–z0–9]*, ….

2.10 Experiments

By following the structural SVM algorithm described in Sect. 3.5, our entity type classifier was implemented by training classification models with the manually annotated data. Figure 6 is a paper's abstract annotated with author defined topic keywords in computer science domain which is used for training structural SVM models.

In our experiments, 10-fold cross evaluation is performed and the average performances are $F_1 = 39.22$ (Korean abstracts) and $F_1 = 38.28$ (English abstracts), respectively, as shown in Table 3. It should be noted that the experiments can be done automatically by executing the prepared scripts which divide into training/testing data, learning structural SVM models, and performing 10-fold cross evaluation based on the trained models. In addition, to show the proposed methods superiority, we performed additional experiments with CRF++, a popular machine learning toolkit based on conditional random field algorithm (2016). In our experiments, the CRF++ based implementations didn't outperform the Struct-SVM based implementations in terms of both accuracy and training time. In case of CRF++, the iteration numbers are different because they are resolved by the soft margin parameter ($C = 4.0$, default setting).

Considering that our framework is in baby step, the performance still promising. However, if we simply address some weaknesses of our current implementation, the amount of training data is not sufficient and our NER module is in immature stage compared with other outstanding NER softwares. To improve the classification performances, devising the current implementation through expression normalization, simplifying entity types, domain dictionary embedding in the classification phrase, and performance optimization with parameter tuning in very essential together with the increasing the number of

<KEYWORD>Naive Bayesian</KEYWORD> learning has been widely used in many <KEYWORD>data mining</KEYWORD> applications, and it performs surprisingly well on many applications. However, due to the assumption that all attributes are equally important in naive Bayesian learning, the posterior probabilities estimated by naive Bayesian are sometimes poor. In this paper, we propose more fine-grained weighting methods, called value weighting, in the context of naive Bayesian learning. While the current weighting methods assign a weight to each attribute, we assign a weight to each attribute value. We investigate how the proposed value weighting effects the performance of naive Bayesian learning. We develop new methods, using <KEYWORD>gradient descent</KEYWORD> method, for both value weighting and feature weighting in the context of naive Bayesian. The performance of the proposed methods has been compared with the attribute weighting method and general Naive bayesian, and the value weighting method showed better in most cases.

Fig. 6 An example of training data

Table 3 Performances according to training algorithms and language

Algorithm (language)	Precision (%)	Recall (%)	F_1 (%)	Training time
Structural SVM (KOR)	39.33	39.11	39.22	24.03 s (400 iterations)
Structural SVM (ENG)	38.23	38.24	38.28	20.05 s (400 iterations)
CRF++ (KOR)	23.29	2.33	4.23	149.82 s (380 iterations)
CRF++ (ENG)	21.05	2.83	4.98	106.37 s (333 iterations)

training data. Those related issues are remained as our future work. Meanwhile, as positive effects, the trained structural SVM models sometimes detect and classify missed candidates that are not identified by human annotators.

3 Applications

To remind extensive usages of our semantic annotation framework in the promising content areas, we introduce three application scenarios where it can play a key role: obtaining technology insights and intelligence, extracting TRIZ theory based deep knowledge, and linking with external data. Our proposed semantic annotation framework can be applied to knowledge extraction modules in each scenario by enhancing extraction performances and extending coverages to other domains.

3.1 Obtaining technology insights and intelligence

Keeping up with rapid advances in scientific researches and patent trends is a challenging task. Nowadays, decision makers in academics, funding agencies, industries, and government agencies not only need to see through latest technology trends, but also be able to predict near future changes in science and technology area. One possible way to deals with the situation to search across large volumes of research articles, patent applications, and patents granted. A recent approach (Dey et al. 2014) presented an automated topical analysis and insight generation from those large heterogeneous text collections of publications and patents. It allows intelligent analysis, such as identifying emerging trends of research, visualizing domain areas' changes over time, building area-wise competition profiles, etc. Meanwhile, their underlying representations are simple n-grams which do not consider associated relations between informative entities. As another example, (Park and Leydesdorff 2013) performed a semantic network analysis on paper titles considering countries to investigate the social and semantic in emerging "big data" research.

Our tag set design and annotation coverage can be extended by allowing representations, such as verb-level expressions that explain factual evidences. Considering the success of IBM watson discovery advisor (2014), we can also take into account of constructing a map of information reasoning over knowledge extracted from scientific publications and real world experiment data. Although constructing flawless information map is still quite far, we can start with aggregating evidence-based knowledge that discusses new inventions, experimental improvements with advanced techniques, new solutions for issue problems, and so on.

3.2 TRIZ theory based deep knowledge extraction

It is already known that the effective use of patent knowledge resources is beneficial for reducing product development time and saving R&D funding. Especially, invention patents containing the highest levels of new knowledge are valuable knowledge resources for products invention. However, non-experts are not easy to understand the patents in spite of detailed, practical texts reflecting the whole process of technology development. In this situation, we can think of the employment of the innovation method, so called TRIZ (theory of the solution of inventive problems) for the classification and deep knowledge extraction patents (Gongchang and Fenghai 2014).

Basically, TRIZ based patents classification is to categorize patent documents automatically by following 40 invention principles (He and Loh 2010). This can be considered as classical text classification problem that can be solved by using support vector machine (SVM) algorithm with appropriate term weighting.

However, deep knowledge acquisition from invention patents is quite different because it requires in-depth understanding about technical background, the process of problem solving in the patent claims, and the content of the invention part. We believe that our semantic annotation framework can be adapted to this kind of knowledge acquisition problem by defining new entity and relation types which can annotate answers for "what is the research object of the patent", "what does the function to achieve", "what are the application results", etc. To this ends, argument zoning (AZ) (Teufel and Moens 2002) which can provide an analysis of the rhetorical progression of the scientific arguments also can be revised for knowledge claims made by inventors of invention patents. Meanwhile, (Jung and Park 2015) attempted a semantic (TRIZ) network analysis using the most frequently occurring 100 keywords from official or authoritative guidelines of open public data that are available through government agencies in South Korea. We believe that our proposed method can make a key role in extending the semantic network analysis by extracting the "problem–solution" patterns from the descriptive texts.

3.3 Linking with external data

The usefulness of the constructed knowledge can be maximized if users can access to the knowledge through publicly known application or public data, and vice versa. As a differentiated approach, PubChem (http://pubchem.ncbi.nlm.nih.gov/) delivers information on the biological activities of small molecules in chemistry domain. For example, by using substance/compound database, it provides links to bioassay description, scientific literature, references, and assay data points. Although those links information is constructed with own purposes, needless to say, it would be beneficial also for other science domains if constructed properly.

Our semantic annotation framework is directly applicable to journal services. In case of researchers who frequently use scientific articles, named entity recognition—a kind of semantic annotation—on the target article, will provide self-explainable contents by allowing linking with external knowledge bases, such as DBpedia and Wikipedia. Let's have a look at a journal service web site (http://www.kpubs.org) which implements simple named entity recognition, a kind of semantic annotation, by using DBpedia spotlight (Daiber et al. 2012). Figure 7 shows that automatically identified named entities in an abstract can be linked with the explanations of DBpedia. If our semantic annotation

framework is trained with Korean journal corpus, the services can be linked with the Korean Wikipedia (http://ko.wikipedia.org/).

In tertiary education aspects, social networks (particularly, from close friends or friends' close friends) can be an educational source because explicit knowledge are transmitted through the connections (Fadul 2014). Linking with the external knowledge bases may give similar effects like having close friends' explicit knowledge. We expect that our proposed semantic annotation framework triggers more extended user experiences together with in-depth analysis of core contents of abstracts.

4 Discussion

4.1 Social-science implication

The social innovation process (Cajaiba-Santana 2014) requires attention to the individual persons rather than domain experts. In addition, it is important to effectively train potential R&D leaders for the success of knowledge-based transition economies (Lee 2013). In this line of context, there is a need to develop process or methodology which can expedite the access of the general public to fruitful scientific resources with low cost. It is more plausible to remove some barriers that dissuade the broad circulation of scientific knowledge with possible tools in the era of open science paradigm (CRF++: Yet Another CRF Toolkit 2016).

Our semantic annotation framework can make the scientific knowledge more accessible with the state-of-the art named entity recognition techniques by extracting clue expressions in text. The extracted expressions can be linked, aggregated, and analyzed further according to various types of efforts to social innovation purposes. Our experiments on chemical engineering and ICT domains showed a possibility. In addition, the suggested

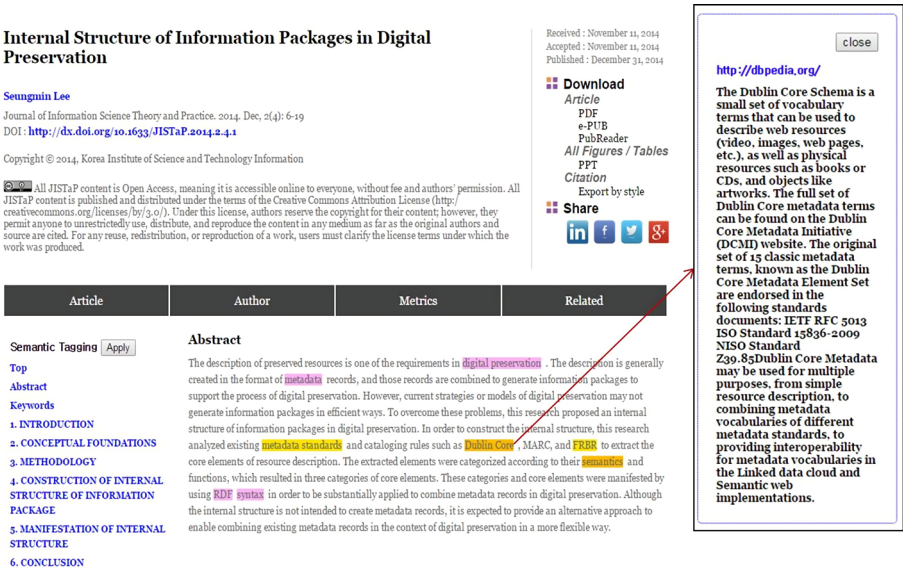

Fig. 7 Examples of linking with external popular data

three scenarios explains possible benefits when our semantic annotation framework is employed successfully in different knowledge-driven applications.

4.2 Additional extensions

To adapt our proposed approaches into other science domains, we need to revise entity types, prepare domain dictionaries with wide coverage, construct sufficient number of annotated data, and optimize named entity recognition (NER) models based on the manually annotated data. It should be noted that our proposed five steps are basic requirements and each step has their own limitations.

Although we defined entity types that can be used in common, they should be revised further according to domain specific characteristics considering users' needs in the target domains. Informational named entity types in math and physics may be different from ours. Moreover, simple entity types may insufficient to annotation meaningful core contents in abstracts. Annotating events or relations should be considered.

4.3 Accuracy considerations

In addition, our semi-automatic annotation tool can be enhanced with extending domain dictionaries including synonyms, acronyms, newly coined technical terms, and multiword expressions in terms of vocabulary coverage. Moreover, to construct annotation data of high quality, we need to allow participations of external experts during revision process because the data quality is closely related with the performance of named entity recognition.

To build a high performing entity recognition models, we need to perform a number of experiments to find out optimal parameters and training features. If lexico-syntactic patterns between different named entity types are similar, trained NER models can hardly classify entities correctly because characteristics of underlying features are not so differential. In addition, there can be huge gaps between trained models with different feature combinations and different parameter settings.

4.4 Methodology aspect

In training a highly performing named entity recognizer, there can be various combinations of popular computational techniques, such as handcrafted rules, maximum entropy (ME), support vector machine (SVM), conditional random field (CRF), etc. Our key concerns in deciding an appropriate machine learning technique are speed and accuracy. Our comparisons with CRF++ have shown that our selected structural SVM method is quite feasible within our purposes. To achieve additional accuracy in named entity recognition task, we can think of a deep neural network (Dos Santos and Guimarães 2014) which is robust against different scales of data, together with minimal turning effort.

5 Conclusion

Techniques that enable automatic extraction of informational named entities existing in scientific articles can help researchers identify information of interest in the growing volume of scientific literature. More specifically, they can contribute substantially to

improve information access of researchers to literature of interest by effectively supporting information retrieval, information extraction and summarization. Although there have been several efforts to extract semantic relations and identify their subordinating entities from scientific articles in English, we report our experimental results both in English and in Korean.

Our semantic annotation framework which includes tag set design, semi-automatic annotation tool, manual annotation data construction, structural SVM based entity recognition, and experiment, is to show one possible avenue towards automatic knowledge generation from scientific publications. The framework presented is capable of scanning millions of scientific literatures, extracting key entities from text, and forecasting meaningful trends and predictions from the identified/extracted entities. In particular, the extracted entities are useful in predicting insight into the lifecycle of emerging technologies, including their maturity, practicality, stages of development, and acceptance by the community.

Social innovation is portrayed as a result of the exchanges of knowledge and resources by actors mobilized through legitimization activities. In this line of context, our framework can be an effective tool which enables communicative actions that are directed towards the achievement of mutual understanding among individuals. We have a firm belief that it can be extended to excavate related scientific knowledge from Korean/English patents as mentioned in our conceptual usage scenarios that may boost a social innovation which can propose new social practices in S&T domains. Moreover, the convergence between research articles and patents will accelerate the open S&T big data paradigm by increasing reusable, trustable knowledge. As our future work, we plan to include relation types, such as APPLY_TO, PROPOSE, RESULT, EVALUTE, COMPARE, etc. in aspects of delineating details of scientific research works.

References

Asahara, M., Matsumoto, Y.: Japanese named entity extraction with redundant morphological analysis. In: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology, vol. 1, pp. 8–15 (2003)

Atdağ, S., Labatut, V.: A comparison of named entity recognition tools applied to biographical texts. In: *ICSCS 2013,* 2nd International conference on systems and computer science, pp. 228–233 (2013)

Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Proceedings of the fifth conference on Applied natural language processing, pp. 194–201 (1998)

Cajaiba-Santana, G.: Social innovation: moving the field forward. A conceptual framework. Technol. Forecast. Soc. Change **82**(1), 42–51 (2014)

Carmel, D., Chang, M.-W., Gabrilovich, E., Hsu, B.P., Wang, K.: ERD'14: Entity Recognition and Disambigutation Challenge. In: SIGIR'14, Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, vol. 48, pp. 1292–1292 (2014)

Chiu, Y., Shih, Y., Lee, Y., Shao, C.: NTUNLP Approaches to Recognizing and Disambiguating Entities in Long and Short Text in the 2014 ERD Challenge. In ERD'14, Proceedings of the first international workshop on Entity recognition & disambiguation, pp. 3–12 (2014)

CRF++: Yet Another CRF Toolkit. https://taku910.github.io/crfpp/ (2016). Accessed 9 June 2016

Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th international conference on semantic systems (I-Semantics), pp. 121–124 (2012)

Dey, L., Mahajan, D., Gupta, H.: Obtaining technology insights from large and heterogeneous document collections. In: 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT), pp. 102–109 (2014)

Dos Santos, C.N., Guimarães, V.: Boosting named entity recognition with neural character embeddings. In: Proceedings of the fifth named entity workshop, joint with 53rd ACL and the 7th IJCNLP, vol. 2014, pp. 25–33 (2015)
Eltyeb, S., Salim, N.: Chemical named entities recognition: a review on approaches and applications. J. Cheminform. **6**(1), 1–12 (2014)
Fadul, J.A.: Big data and knowledge generation in tertiary education in the Philippines. J. Contemp. East. Asia **13**(1), 5–18 (2014)
Ferragina, P., Scaiella, U.: TAGME: One-the-fly Annotation of Short Text Fragmetns (by Wikiepdia Entities). In: CIKM'10, Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1625–1628 (2010)
Gongchang, R., Qi, L., Fenghai, Y.: On classification and extraction of deep knowledge in patents based on TRIZ theory. In: 2014 Fifth international conference on intelligent systems design and engineering applications, pp. 666–670 (2014)
Grishman, R., Borthwick, A., Sterling, J., Agichtein, E.: NYU: description of the MENE named entity system as used in MUC-7. In: Proceedings of the seventh message understanding conference (MUC-7) (1998)
Guo, Y., Korhonen, A., Poibeau, T.: A weakly-supervised approach to argumentative zoning of scientific documents. In: EMNLP'11, Proceedings of the conference on empirical methods in natural language processing, pp. 273–283 (2011)
Gupta, S., Manning, C.: Analyzing the dynamics of research by extracting key aspects of scientific papers. In: Proceedings of 5th international joint conference on natural language processing, pp. 1–9 (2011)
He, C., Loh, H.T.: Pattern-oriented associative rule-based patent classification. Expert Syst. Appl. **37**(3), 2395–2404 (2010)
Ibekwe-SanJuan, F.: Semantic metadata annotation: tagging Medline abstracts for enhanced information access. Aslib Proc. **62**, 476–488 (2010)
IBM Watson Discovery Advisorhttp://www.ibm.com/smarterplanet/us/en/ibmwatson/discovery-advisor.html (2014). Accessed 9 June 2016
Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural SVMs. Mach. Learn. **77**, 27–59 (2009)
Jung, K., Park, H.W.: A semantic (TRIZ) network analysis of South Korea's 'Open Public Data' policy. Gov. Inf. Q. **32**(3), 353–358 (2015)
Lee, Y.-G.: Multidisciplinary Team Research as an Innovation Engine in Knowledge-Based Transition Economies and Implication for Asian Countries -From the Perspective of the Science of Team Science. J. Contemp. East. Asia **12**(1), 49–63 (2013)
Lee, C., Jang, M.G.: A modified fixed-threshold SMO for 1-slack structural SVMs. ETRI J. **32**(1), 120–128 (2010)
Lee, C., Hwang, Y.-G., Oh, H.-J., Lim, S., Heo, J., Lee, C.-H., Kim, H.-J., Wang, J.-H., Jang, M.-G.: Fine-grained named entity recognition using conditional random fields for question answering. In: *Information Retrieval Technology*, vol. 5839, pp. 581–587 (2006)
Lee, C., Ryu, P.-M., Kim, H.: Named entity recognition using a modified Pegasos algorithm. In: CIKM'11, Proceedings of the 20th ACM international conference on information and knowledge management, pp. 2337–2340 (2011)
McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: HLT-NAACL 2003, Proceedings of the seventh conference on natural language learning, vol. 4, pp. 188–191 (2003)
Mizuta, N.C.Y., Korhonen, A., Mullen, T.: Zone analysis in biology articles as a basis for in- formation extraction. Int. J. Med. Informatics. **75**(6), 468–487 (2006)
Murphy, T., Mcintosh, T., Curran, J. R.: Named entity recognition for astronomy literature. In: Proceedings of the Australasian language technology workshop (ALTW), pp. 59–66 (2006)
Open Sciencehttps://en.wikipedia.org/wiki/Open_science (2016). Accessed 9 June 2016
Park, H.W., Leydesdorff, L.: Decomposing social and semantic networks in emerging 'big data' research. J. Informetr. **7**(3), 756–765 (2013)
Park, Y.M., Kang, S.W., Seo, J.G.: Title named entity recognition using Wikipedia and abbreviation generation. In: International conference on big data and smart computing (BIGCOMP), pp. 169–172 (2014)
Phillips, F.: Triple helix and the circle of innovation. J. Contemp. East. Asia **13**(1), 57–68 (2013)
Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: 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 (EACL), pp. 102–107 (2012)

Tateisi, Y., Shidahara, Y., Miyao, Y., Aizawa, A.: Annotation of computer science papers for semantic relation extraction. In: *Proceedings of the 9th international conference on language resources and evaluation*, pp. 1423–1429 (2014)

Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Comput. Linguist. **28**, 409–445 (2002)

Teufel, S., Batchelor, C.: Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 conference on empirical methods in natural language processing, pp. 1493–1502 (2009)

Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML'04 Proceedings of the twenty-first international conference on Machine learning, p. 104 (2004)

Yoon, B., Park, I., Coh, B.Y.: Exploring technological opportunities by linking technology and products: application of morphology analysis and text mining. Technol. Forecast. Soc. Change **86**, 287–303 (2014)