Journal of Informetrics 9 (2015) 455–465

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Identifying entities from scientific publications: A comparison of vocabulary- and model-based methods

Erjia Yan*, Yongjun Zhu

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 22 April 2015
Accepted 22 April 2015
Available online 16 May 2015

Keywords:
Entity extraction
Vocabulary
Dictionary
Conditional random fields
Content-aware

ABSTRACT

The objective of this study is to evaluate the performance of five entity extraction methods for the task of identifying entities from scientific publications, including two vocabulary-based methods (a keyword-based and a Wikipedia-based) and three model-based methods (conditional random fields (CRF), CRF with keyword-based dictionary, and CRF with Wikipedia-based dictionary). These methods are applied to an annotated test set of publications in computer science. Precision, recall, accuracy, area under the ROC curve, and area under the precision-recall curve are employed as the evaluative indicators. Results show that the model-based methods outperform the vocabulary-based ones, among which CRF with keyword-based dictionary has the best performance. Between the two vocabulary-based methods, the keyword-based one has a higher recall and the Wikipedia-based one has a higher precision. The findings of this study help inform the understanding of informetric research at a more granular level.

1. Introduction

Publication data embody the very essence of humans' scientific and technological advances. These data have been continuously examined through multidisciplinary efforts. Citation-based methods have traditionally been employed to assess research impact; modern statistical methods have employed various bibliometric networks to cluster research specialties, detect author communities, and identify research topics. While these efforts have revealed patterns of scholarly communication, elucidated the scientific workforce, determined mechanisms of impact assessment, and addressed a slew of issues related to disciplinarity and interdisciplinarity, they were largely driven by the analysis of existing publication metadata (e.g., authors, titles, journals, and references). Consequently, we have limited understanding of the ways to analyze the content of individual papers. Moreover, because knowledge is more effectively expressed through unstructured or semi-structured contents, such as titles, abstracts, keywords, or even full-text, we have yet to find out how to use the contents to examine knowledge production and innovation. Therefore, the current study intends to tackle the complexity of unstructured and semi-structured contents, with a focus on detecting entities – expressions in the contents that convey research-relevant information – from texts. This study is part of a larger effort to understand the mechanisms of innovation-making through content-aware approaches.

Entity extraction is not a new idea: it is an important sub-task of information extraction and is sometime referred to as named entity extraction and classification (NERC) (Nadeau & Sekine, 2007). The goal of NERC is to identify and classify

* Corresponding author. Tel.: +1 215 895 1459; fax: +1 215 895 2494.
E-mail addresses: ey86@drexel.edu (E. Yan), yz493@drexel.edu (Y. Zhu).

http://dx.doi.org/10.1016/j.joi.2015.04.003
1751-1577/© 2015 Elsevier Ltd. All rights reserved.

named entities from large, heterogeneous text corpora. Names, as Kripke (1972) puts it, are "rigid designators" (p. 48). Thus, earlier NERC tasks were largely focused on the extraction of proper names from texts (Thielen, 1995), such as the names of locations, people, and organizations—collectively known as "enamex"; data and time types ("timex") and money and percent types ("numex") were also recognized as entity types for NERC tasks (Nadeau & Sekine, 2007). Recent advances in bioinformatics has also incorporated the identification of biomedical entities, such as genes, compounds, drugs, proteins, and diseases, into the NERC framework (e.g., Bekhuis, 2006; Jensen, Saric, & Bork, 2006; Swanson, Smalheiser, & Torvik, 2006). Meanwhile, we acknowledge the fact that NERC is not restricted to academic research—there are successful commercialized NERC systems for large synchronized language analyses, particularly for defense applications. For instance, the U.S. Defense Advanced Research Projects Agency (DARPA) has allocated more than $100 million between 2003 and 2005 for projects on Automated Speech and Text Exploitation in Multiple Languages (DARPA, 2005). Traditionally, domain specific dictionaries were employed to extract named entities from texts; however, this technique did not scale up with the emergence of new named entities and its performance is impaired by the fuzziness of the natural language (Sekine & Nobata, 2004). Modern statistical methods, on the other hand, are capable of recognizing and disambiguating new named entities, through supervised methods, such as hidden Markov models (HMM; Bikel, Miller, Schwartz, & Weischedel, 1997) and conditional random fields (CRF; Lafferty, McCallum, & Pereira, 2001) or semi- or unsupervised methods, such as bootstrapping (Riloff & Jones, 1999). These methods will be surveyed in the literature review section.

These statistical methods have been applied to extract "enamex", "timex", "numex", and biomedical-related named entities and high precision and recall have been reported (e.g., Collier, Nobata, & Tsujii, 2000; Torii, Hu, Wu, & Liu, 2009; Jiang et al., 2011). However, as Nadeau and Sekine (2007) argued, "[t]he impact of textual genre...and domain...has been rather neglected in the NERC literature...[f]ew studies are specifically devoted to diverse genres and domains" (p.2). Since then, there have been attempts to extend the scope of NERC by extracting entities from scientific literature (e.g., He & Kayaalp, 2008; Prokofyev, Demartini, & Cudré-Mauroux, 2014). Thus, this study is motivated to develop this body of literature by evaluating the performance of several entity extraction methods on a text corpus that contains scientific publications. This textual genre, as a distinctive science communication channel, exhibits its own discourse-related characteristics (Hyland, 2000; Demarest and Sugimoto, in press). Papers in five leading computer science journals are selected as the data set. Several vocabulary- and statistical model-based methods are employed to identify entities from this data set, including two vocabulary-based methods (i.e., a keyword-based and a Wikipedia-based) and three model-based methods (i.e., CRF, CRF with keyword-based dictionary, and CRF with Wikipedia-based dictionary approaches).

Findings from this study will advance the methods of scholarly data mining as well as the application of these methods for content-aware studies of knowledge production and innovation. Conducting content-aware research has the readily apparent advantage of gaining explicit and fine-grained perspectives of how different entities are embedded and related. It will also enhance our understanding of the provenance of knowledge as codified by entities. Results of this research will lay a foundation for these efforts and help inform scientists and scholars for more granular analyses of the history, contemporary landscape, and future trajectories of domains.

2. Related work

This section reviews several types of entity extraction methods, including vocabulary-based, semi- or unsupervised methods, and supervised methods.

2.1. Vocabulary-based methods

Vocabulary-based methods have been employed to identify and disambiguate the concepts of interest, such as title words (e.g., Swanson, 1986), subject headings (e.g., Swanson et al., 2006) and thesaurus dictionaries (e.g., Ding, Chowdhury, & Foo, 2001; Lou & Qiu, 2014). A pioneering study by Swanson (1986) has built off title word co-occurrence relations to detect latent entity relations. Different from finding co-occurrence relations between two directly connected entities, Swanson's approach used two disjoint sets of records and identified a list of terms that co-occurred with both sets. This approach has been empirically tested and it has helped verify some previously overlooked relations such as fish oil and Raynaud's syndrome, magnesium and migraine, somatomedin C and arginine, and even viruses as weapons, according to a review article by Bekhuis (2006). It has been suggested that the use of controlled vocabularies can reduce the ambiguity of the natural language (e.g., Swanson et al., 2006). In biomedical domains, the Medical Subject Headings (MeSH) has been widely used to retrieve medical publications (e.g., Lowe & Barnett, 1994) and to find the relatedness of medical terms (e.g., Nelson, Johnston, & Humphreys, 2001). For instance, Swanson's approach was improved by the use of MeSH terms to enhance its efficiency (Swanson et al., 2006).

Despite the effort of controlled vocabularies such as MeSH to consistently index bio-entities, it was found that they may not fully address the nomenclature problems of synonyms, noun phrases, and acronyms (Morgan, Hirschman, Colosimo, Yeh, & Colombe, 2004; Galvez & de Moya-Anegón, 2012). Thus, more specialized dictionaries and ontologies have been designed and experimented with, serving the goal to discriminate and integrate genes, compounds, drugs, proteins, and diseases in various orthographic forms (e.g., Humphreys, Lindberg, Schoolman, & Barnett, 1998; Ashburner et al., 2000; Jensen et al., 2006; Liu, Hu, Torii, Wu, & Friedman, 2006; Frijters et al., 2008, 2010; Galvez & de Moya-Anegón, 2012). Among these, the Unified Medical Language System (UMLS; Humphreys et al., 1998) and Gene Ontology (GO; Ashburner et al.,

2000) are probably the most representative. These ontologies were designed to address "the lack of a standard language in medicine" (, p. 2). In this regard, they represent prior knowledge by incorporating a variety of existing thesauri and lexicons. They also provide hierarchies, semantic types, and semantic relations to existing controlled vocabularies to support the annotation of gene products, drug discovery and repurposing, instance sequences, or other biomedical tasks (e.g., Chagoyen, Carmona-Saez, Shatkay, Carazo, & Pascual-Montano, 2006; Andronis, Sharma, Virvilis, Deftereos, & Persidis, 2011; Galvez & de Moya-Anegón, 2012). These vocabulary-based methods are dependent on the prior knowledge represented by dictionaries or ontologies. They thus lack the capacity to identify new concepts or ideas that are not readily indexed. This drawback is manifested in fast-growing fields where new terms are emerging rapidly or interdisciplinary fields where domain specific dictionaries may no longer be effective (Demner-Fushman, Chapman, & McDonald, 2009).

2.2. *Semi- or unsupervised methods*

The semi-supervised methods for entity extraction typically refer to the bootstrapping technique (Riloff & Jones, 1999; Riloff, Wiebe, & Wilson, 2003). It is a self-sustaining technique used to iteratively improve a classifier's performance. The first step is to select seed terms; previous studies have shown that a small number of seed terms are sufficient to produce high-level relevance (Riloff & Jones, 1999). The method then learns the contextual patterns of seed terms and uses the learned patterns to select new terms. This iterative method is also known as meta-bootstrapping (Riloff & Jones, 1999). The calculation of pattern scores and entity scores determine the effectiveness of the bootstrapping method. If a pattern gets a higher score, then it is selected into the candidate pattern pool. Entities extracted by these candidate patterns are considered as candidate entities. Heuristic rules have been used to generate patterns (Thelen & Riloff, 2002), for instance, the presence of uppercase letters, sentence structure, and the collocation of words. Bootstrapping can have a binary classifier (i.e., is or is not a named entity) or a multi-class classifier in the case of labeled data. It has been demonstrated that "predicting the labels of unlabeled entities" can improve the performance of the bootstrapping method (Gupta & Manning, 2014a, p. 9). Yet, this may require some domain dictionaries to estimate the labels of the extracted entities. Overall, bootstrapping is an "effective, interpretable" (Gupta & Manning, 2014a, p. 1) method and previous studies have shown that this semi-supervised learning method performed well on extracting entities from subject areas such as terrorism (Thelen & Riloff, 2002), law (Nallapati & Manning, 2008), and medicine (Gupta & Manning, 2014b).

Unsupervised methods use lexical resources (e.g., WordNet or web queries) and lexical patterns (e.g., the "such as" pattern) to extract named entities (Nadeau & Sekine, 2007). For instance, Alfonseca and Manandhar (2002) mapped terms through WordNet synset to assess the likelihood that terms be assigned to certain classes. Likewise, using web queries as the lexical resources, Etzioni et al. (2005) estimated the dependency between two expressions and used the dependency probabilities to assign entities to classes—the system was known as KnowItAll. KnowItAll required high volumes of web queries and the extraction process had to be updated every time a new relation was introduced. This scalability issue was resolved by the Open Information Extraction project TextRunner (Banko, Cafarella, Soderland, Broadhead, & Etzioni, 2007), a highly efficient entity extraction system that utilized web data. TextRunner opened up a new paradigm for entity extraction and there are ongoing efforts to improve this system (e.g., Christensen, Soderland, & Etzioni, 2010; Akbik & Löser, 2012; Del Corro & Gemulla, 2013). Studies that exploited lexical patterns include Collins and Singer's (1999) work on unsupervised models for named entity classification. The model relied on spelling and contextual rules—spelling rules applied to target strings whereas contextual rules focused on the surrounding terms of the target strings. Built on Collins and Singer's (1999) work, Nadeau, Turney, and Matwin (2006) improved the unsupervised method through two ways: one was to add a named-entity disambiguation component which is an advancement to earlier studies (e.g., Etzioni et al., 2005) that were largely predicated on entity extraction but not recognition or disambiguation; and the other way was to utilize text without standard grammatical structures, such as tables and lists, in HTML documents. The merit of the semi- and unsupervised methods is that they require limited human intervention; however, to ensure satisfying performances, external data sources or heuristic rules are needed, thus posing challenges toward a streamlined and transparent research design.

2.3. *Supervised methods*

The supervised entity extraction models primarily include maximum entropy models (e.g., Borthwick, Sterling, Agichtein, & Grishman, 1998), support vector machines (e.g., Krauthammer & Nenadic, 2004; Takeuchi & Collier, 2005), decision trees (e.g., Prokofyev et al., 2014), hidden Markov models (HMM; e.g., Collier et al., 2000), and conditional random fields (CRF; e.g., Torii et al., 2009; Jiang et al., 2011). These methods are typically applied to extract named entities from non-scientific documents such as news articles. A recent work (Prokofyev et al., 2014) on extracting entities from scientific publications is perhaps the most similar to the current study. Publications on information retrieval and physics were included as the data set and a supervised decision tree model was applied: it is found that the use of external knowledge bases, such as DBLP and DBPedia, has improved the performance of the decision tree-based models. Both HMM and CRF belong to sequence models and are capable of modeling interdependent variables. This is an essential feature because words in a sentence are dependent (Sutton & McCallum, 2006). The difference between HMM and CRF is that HMM is generative and is based on a model of joint distributions whereas CRF is discriminative and is based on a model of conditional distributions (Lafferty et al., 2001; Sutton & McCallum, 2006). While both models have merit, it is suggested that CRF is more flexible in data fitting and is "better suited to including rich, overlapping features" (Sutton & McCallum, 2006, p. 8). CRF is therefore employed in this study and its performance on extracting entities from publications in computer science is assessed.

3. Data

Publication data were gathered from five flagship computer science journals. We chose the *Journal of the ACM (JACM)* as the seed journal. Top five most cited journals by *JACM* were obtained using journal citation data from the Web of Science database. For each of the five journals, their top five most cited journals were identified, resulting in 25 cited journal instances. Next, aggregating these instances, we then obtained a list of the most cited journals by the five journals. If the top five journals are the same five journals that *JACM* cited the most, then these five journals are our core journals of analysis; otherwise, repeat the above two steps until a stable set of five journals is reached. This approach of determining core journals was pioneered by Hirst (1978). The resulted five journals are: *Journal of the ACM*, *Theoretical Computer Science*, *SIAM Journal on Computing*, *Information and Computation*, and *Journal of Computer and System Sciences*.

Publications between 2004 and 2014 (March) were downloaded as the data set of this study. The total number of papers is 7262. To assess the performance of the entity extraction methods, two sets were prepared: a training set and a test set. To prepare the training set, five papers were randomly sampled for each year between 2004 and 2014 (total number of papers: 55; total number of title/abstract terms: 9022). This set will be used to train the supervised CRF model. This stratified sampling method was also used to sample publications for the test set. Five papers were randomly sampled for each year between 2004 and 2014 (total number of papers: 55; total number of title/abstract terms: 10,283). The training data set for "enamex"-based entity extractions is typically larger: for instance, more than 200 thousand words from 946 documents were used to train the CRF model to learn labels on persons, locations, organizations, and miscellaneous entities (McCallum, 2002; Sutton & McCallum, 2011). The size of the training data may relate to the performance of the CRF model and we see it as a future research direction to explore the impact of the size and disciplinary identities on the performance of the CRF model. For the training set, two coders independently annotated the titles and abstracts using a binary rule; that is, 0 for non-entities and 1 for entities. For the first run, an inter-rater reliability (IRR) of 0.55 was attained (based on Cohen's kappa). Coders compared the annotations, identified and discussed the disagreement, and set up more explicit coding rules and another independent annotation was implemented. This run yielded an IRR of 0.75, a satisfying agreement for categorical data analysis (e.g., Hallgren, 2012). One coder then annotated the 55 papers in the test set using the same standard. We consider the resulted annotated set as the rubric to benchmark the performance of all methods to be evaluated in this study.

4. Methods

4.1. Procedures

We propose and evaluate five methods to extract entities from publications in computer science.

1. Keyword-based: the first method constructed a keyword dictionary from publications in the data set. The keyword dictionary contains terms from both author provided keywords and database provided descriptors (size of the dictionary: 21,349 terms). To ensure an accurate term matching, a plural-singular transformation was preferred over lemmatization because the latter approach may change the content of entities, for instance, from "nearest neighbors" to "near neighbor". For plural-singular transformation, if the plural form of a keyword was used, its singular and plural forms were added to the dictionary (e.g., if "nearest neighbors" is a keyword, then both "nearest neighbors" and "nearest neighbor" are added to the dictionary). If no plural form of a keyword was present, only its singular form was added to the dictionary (e.g., if "multi-word synchronization" is the only form, then it is directly added to the dictionary without the need to add its plural form). This method requires the least processing and is adopted as the baseline approach to assess the performance of other more advanced approaches.
2. Wikipedia-based: similar to the first method, a dictionary was constructed using terms from Wikipedia entries. To collect domain specific terms, we started off from the computer science category in Wikipedia (http://en.wikipedia.org/wiki/Category:Computer_science). This category contains 14 first-level subcategories and 19 first-level pages. As we randomly sampled some of the pages at each level, pages have become less relevant to computer science. To make a proper balance between the size and quality of the dictionary, we included all terms at level five and the same plural-singular transformation was performed (size of the dictionary: 49,993 terms).
3. The conditional random fields (CRF) method: the third method used the annotated training set to train the CRF model. The trained CRF model will be applied to the test set to extract entities.
4. CRF with Keyword-based dictionary: similar to the third method, it used the annotated training set for the supervised learning. Additionally, a keyword dictionary (developed in the first method) was incorporated in the gazette function of the CRF model with the goal to improve its performance (Kim, Ohta, Tsuruoka, Tateisi, & Collier, 2004; Sutton & McCallum, 2006; Pawar, Srivastava, & Palshikar, 2012).
5. CRF with Wikipedia-based dictionary: similar to the fourth method that used a keyword-based dictionary as the gazette, this method used a Wikipedia-based dictionary (developed in the second method) as the gazette in the CRF model.

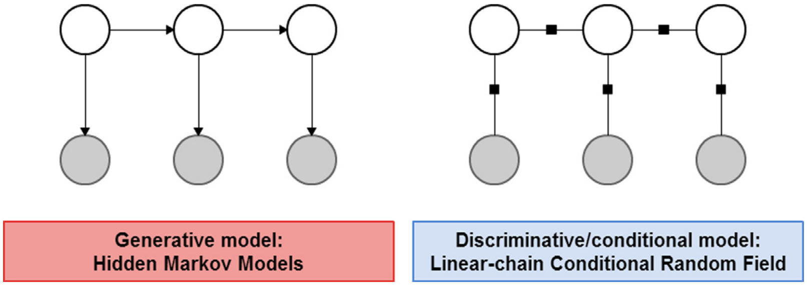

Generative model:
Hidden Markov Models
Discriminative/conditional model:
Linear-chain Conditional Random Field

Fig. 1. Two sequence models: HMM and CRF (Sutton & McCallum, 2006).

Table 1
Test outcomes.

Method	TN	TP	FN	FP	TN + TP
Keyword	7779	793	441	1270	8572
Wikipedia	8819	215	1019	230	9034
CRF	8461	787	447	588	9248
CRF + keyword-based dictionary	8406	840	394	643	9246
CRF + Wikipedia-based dictionary	8464	811	423	585	9275

Table 2
Precision, recall, AUC, and AUP of the five methods.

Method	Precision	Recall	ACC	AUC	AUP
Keyword	0.38	0.64	0.83	0.75	0.30
Wikipedia	0.48	0.17	0.88	0.57	0.20
CRF	0.57	0.64	0.90	0.75	0.42
CRF + keyword-based dictionary	0.57	0.68	0.90	0.80	0.44
CRF + Wikipedia-based dictionary	0.58	0.66	0.90	0.79	0.43

A standard CRF model supports the definition of multiple classes, such as locations, organizations, and people (Sutton & McCallum, 2006) or problem, treat, and test (Jiang et al., 2011). As for publications, however, the design of a multi-class model may meet with several challenges, for instance, disciplinary barriers, language ambiguities, a lack of consensus on class definitions, and a lack of rich contexts. Therefore, a one-class model was adopted in this study. A multi-class model will be endeavored as future work.

4.2. Conditional random fields

This section briefly introduces the employed CRF model; detailed mathematical presentations can be found in Appendix A. CRF is a sequence model based on the idea of obtaining output probabilities using the input sequence. Sequence model is first trained to learn some sequence observations. It then provides a predictive measure of similarity to other sequences as an output for any new sequence observations (Sutton & McCallum, 2006). Sequence model is thus an ideal platform to model natural language patterns because words in sentences form a sequential flow of semantics.

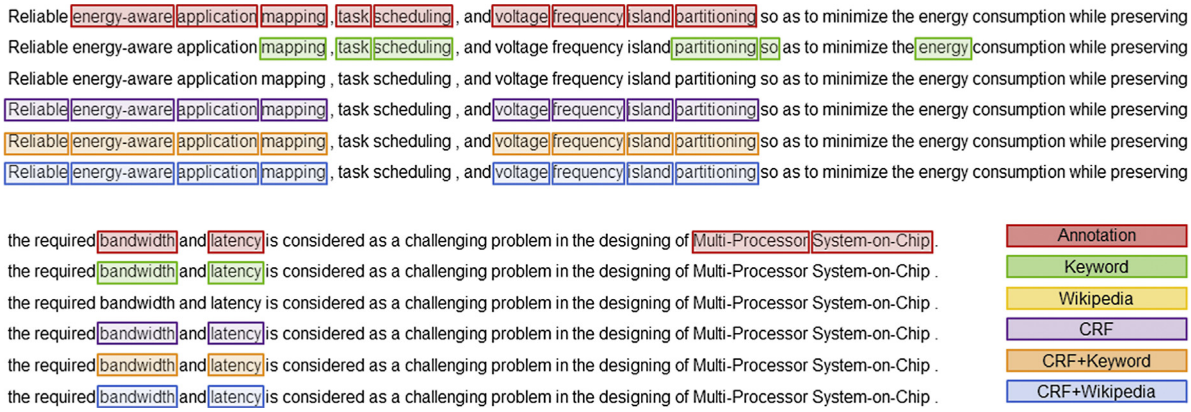
There are two representative sequence models (Fig. 1): one is CRF and the other is the HMM. HMM is a generative model based on joint distributions ($p(y,x)$), and CRF is a discriminative model based on conditional distributions ($p(y|x)$) (Lafferty et al., 2001; Sutton & McCallum, 2006). Therefore, CRF does not need to include a model of $p(x)$ and is more flexible than HMM (Sutton & McCallum, 2006; de Souza, Pizzolato, & dos Santos Anjo, 2012).

5. Results

The five methods are evaluated using several approaches: Section 5.1 reports raw test outcomes (Table 1) as well as precision, recall, accuracy (ACC), area under the ROC curve (AUC), and area under the precision-recall curve (AUP) (Table 2); Section 5.2 uses the Mann–Whitney test of asymptotic significance to assess the between-group differences of the five methods; Section 5.3 illustrates ROC and precision-recall curves to give visual presentations of the results.

5.1. Performance of the five methods

Table 1 presents the test outcomes on true negative (TN), true positive (TP), false negative (FN), and false positive (FP) scores.



Reliable energy-aware application mapping , task scheduling , and voltage frequency island partitioning so as to minimize the energy consumption while preserving

Reliable energy-aware application mapping , task scheduling , and voltage frequency island partitioning so as to minimize the energy consumption while preserving

Reliable energy-aware application mapping , task scheduling , and voltage frequency island partitioning so as to minimize the energy consumption while preserving

Reliable energy-aware application mapping , task scheduling , and voltage frequency island partitioning so as to minimize the energy consumption while preserving

Reliable energy-aware application mapping , task scheduling , and voltage frequency island partitioning so as to minimize the energy consumption while preserving

Reliable energy-aware application mapping , task scheduling , and voltage frequency island partitioning so as to minimize the energy consumption while preserving

the required bandwidth and latency is considered as a challenging problem in the designing of Multi-Processor System-on-Chip .

the required bandwidth and latency is considered as a challenging problem in the designing of Multi-Processor System-on-Chip .

the required bandwidth and latency is considered as a challenging problem in the designing of Multi-Processor System-on-Chip .

the required bandwidth and latency is considered as a challenging problem in the designing of Multi-Processor System-on-Chip .

the required bandwidth and latency is considered as a challenging problem in the designing of Multi-Processor System-on-Chip .

the required bandwidth and latency is considered as a challenging problem in the designing of Multi-Processor System-on-Chip .

Fig. 2. An example of entity extraction using the five methods.

Table 1 shows that based on TN + TP, CRF with Wikipedia-based dictionary has the highest score: it made correct judgments on 9275 words out of 10,283 in total. While the Wikipedia-based approach had the highest true negative score and the lowest false positive score, it resulted in the smallest true positive score and the highest false negative score, suggesting that it tended to be rather stringent in determine entities and thus missed a considerably amount of entities. The keyword-based approach, on the other hand, resulted in the highest false positive score, indicating that it tended to be a relaxed method to extract entities, however, at a low accuracy (its true positive ranked the third). For the three CRF-based methods, the standard CRF and CRF with Wikipedia-based dictionary had a similar performance; CRF with keyword-based dictionary had a higher true positive score but also a higher false positive score.

Results on precision, recall, ACC, AUC, and AUP are reported in Table 2. Precision is calculated by $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$, recall is $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, and ACC is $\frac{\mathrm{TN+TP}}{\mathrm{TN+TP+FN+FP}}$.

Table 2 shows that the two vocabulary-based methods have the lowest precision. In regards to recall, the keyword-based approach has yielded considerably higher recall than the Wikipedia-based approach, while Wikipedia-based approach has recorded higher precision. The higher precision of the Wikipedia-based approach suggests that it has identified key terms more precisely than the keyword-based approach. Meanwhile, the low recall indicates that the Wikipedia-based dictionary did not include as many relevant terms as the keyword-based dictionary. The way we constructed the Wikipedia-based dictionary is to include terms in the Wikipedia category on computer science. A possible issue may be caused by the varied depths of different subcategories—while level five may be sufficient for some subcategories, it may be inadequate (i.e., not enough relevant terms were included) or overfitting (i.e., a higher degree of noise were included) for some other subcategories. This issue could be solved by developing rule-based methods to systematically extract more key terms while reducing the noise.

Three CRF-based methods have improved both precision and recall from the two vocabulary-based methods, among which CRF with Wikipedia-based dictionary has the highest precision at the 0.58 level and CRF with keyword-based dictionary has the highest recall at the 0.68 level. The results demonstrate that the statistical training through CRF has improved the performance of the entity extraction task. Prior research has found a satisfying performance of CRF in recognizing named entities, such as locations, names, and organizations (e.g., Nadeau & Sekine, 2007; Sutton & McCallum, 2006). This study adds new findings to this area by revealing the value of CRF in identifying key terms from publications in computer science. In the meantime, we also noticed a gap between the performances of entity recognitions for non-scientific publications and scientific publications: for instance, a precision of 0.89 was reported for a company name extraction task (Tang, Hong, Li, & Liang, 2006) and whereas a precision at the 0.6 level was reported in this study as well as an earlier scientific entity extraction task on biological entities (He & Kayaalp, 2008). This may be attributed to the fact that “enamex” type entities are typically associated with more fixed locations in sentences such that it makes easier for the CRF model to learn these structural patterns; entities in publications, on the other hand, do not necessarily possess such patterns as they can occur in both expected and unexpected locations in sentences. Furthermore, the size of the training data may also be a contributing factor: the training set was limited to 55 papers with 10,283 title/abstract terms whereas the training sets for “enamex”, “timex”, and “numex” extractions are much larger in size (McCallum, 2002; Nadeau & Sekine, 2007; Sutton & McCallum, 2011).

5.2. Relationships of the five methods

To begin with, we use an excerpted sentence from the data set to exemplify the extraction results of the five methods (Fig. 2).

The keyword-based method has extracted entities including “mapping”, “task scheduling”, “partitioning”, and “energy”—not surprisingly because these are all common keywords in computer science. It, however, also extracted “so”

Table 3
Mann–Whitney test of asymptotic significance (2-tailed).

	Annotation	Keyword	Wikipedia	CRF	CRF + keyword	CRF + Wiki
Annotation	1					
Keyword	0.000	1				
Wikipedia	0.000	0.000	1			
CRF	0.003	0.000	0.000	1		
CRF + keyword	0.000	0.000	0.000	0.029	1	
CRF + Wiki	0.001	0.000	0.000	0.668	0.080	1

which might mean a UNIX shared library format file (e.g., "abc.so"); meanwhile, it missed entities such as "voltage frequency island" because these terms were not included in the keyword dictionary. The Wikipedia-based method did not extract any entity from the excerpted sentence, because none of the words were included in the Wikipedia dictionary used in this study. The use of the full Wikipedia entry dictionary is likely to capture entities from this sentence, but irrelevant entities may also be extracted as a consequence because of the low precision of the exact Wikipedia matching (Prokofyev et al., 2014). The three CRF-based methods extracted the same set of entities from the sentence: it missed the entity "task scheduling", "Multi-Processor", and "System-on-Chip" and mislabeled "Reliable" which is seen as a describing term for an entity.

To further assess the differences among the five methods, we used the Mann–Whitney test of asymptotic significance and applied it to the whole test set. It is an effective non-parametric significance test of between-group differences. The results are reported in Table 3; in particular, the human annotated results were also included to evaluate the difference between the human annotation and the five methods.

Table 3 shows that the human annotation is statistically different from any of the five methods; results of CRF and CRF with Wikipedia-based dictionary are probably the closest to the human annotation, as indicated by the p value, though still smaller than 0.05. Two vocabulary-based methods generated results different from the human annotation and also from any of the CRF-based methods. In the meantime, CRF with Wikipedia-based dictionary did not deliver results that were statistically different from CRF or CRF with keyword-based dictionary ($p > 0.05$)—particularly that the p value is much larger than 0.05 for the test of asymptotic significance between CRF and CRF with Wikipedia-based dictionary ($p = 0.668$). Although CRF and CRF with keyword-based dictionary yielded statistically different results, the p value is close to the α-level of 0.05. The test of asymptotic significance complements the performance tests in Tables 1 and 2: it shows that, first, there are still discrepancies between the results obtained from the proposed methods and the "gold standard" in the form of human annotation; second, vocabulary-based methods performed differently from each other as well as from the model-based methods; third, the CRF-based methods yielded more similar results and, in particular, the standard CRF may replicate the results of CRF with Wikipedia-based dictionary.

5.3. Visualizing the results through AUC and AUP

To gain an understanding of the relationship between precision and recall, we employed ROC and precision-recall curves (Figs. 3 and 4).

We see from the ROC curves that the Wikipedia-based approach behaved differently from the other four methods—its curve formed a jagged diagonal line which exemplifies an AUC of 0.57. The other four methods, including the keyword- and three CRF-based approaches, exhibited more similar distribution patterns. They had small false positive rates (FPR) when the true positive rates (TPR) stayed within 0.7 and their FPR hiked when the TPR was above 0.7. Thus, a TPR of 0.7 is a noticeable upper threshold for the four methods under study. In terms of the precision-recall curves, the Wikipedia-based approach maintained a precision of 0.5 when the recall was less than 0.2 and its precision plunged as the recall went higher. The tipping point for the keyword-based method was 0.6 in recall. The precision-recall curves for the three CRF-based methods displayed similar distribution patterns in that they maintained the precision around 0.6 when the recall was smaller than 0.7. Their precision dropped as the recall went up—CRF was the first to drop, followed by CRF with Wikipedia-based dictionary and CRF with keyword-based dictionary.

Last, to obtain a clearer insights on the distributions of FPR, TFR, precision, and recall, we used α-binominal model to smooth the ROC curves and precision-recall curves (Brodersen, Ong, Stephan, & Buhmann, 2010), as shown in Fig. 4. Both images illustrate that CRF with keyword-based dictionary and CRF with Wikipedia-based dictionary are the best performing approaches, following by the standard CRF and the two vocabulary-based methods. Between the two vocabulary-based methods, assessed through ROC curves, the keyword-based method had an overarching better performance than the Wikipedia-based one, while in the precision-recall curves, the Wikipedia-based method started off with a higher precision and was surpassed by the keyword-based method when recall went higher than 0.3. These findings are consistent with the ones obtained from the non-smoothed curves in Fig. 3 as well as precision, recall, AUC, and AUP in Table 2.

6. Discussion and conclusion

This paper has employed and evaluated five approaches in extracting entities from publications in computer science, including a keyword-based method, a Wikipedia-based method, CRF, CRF with keyword-based dictionary, and CRF with

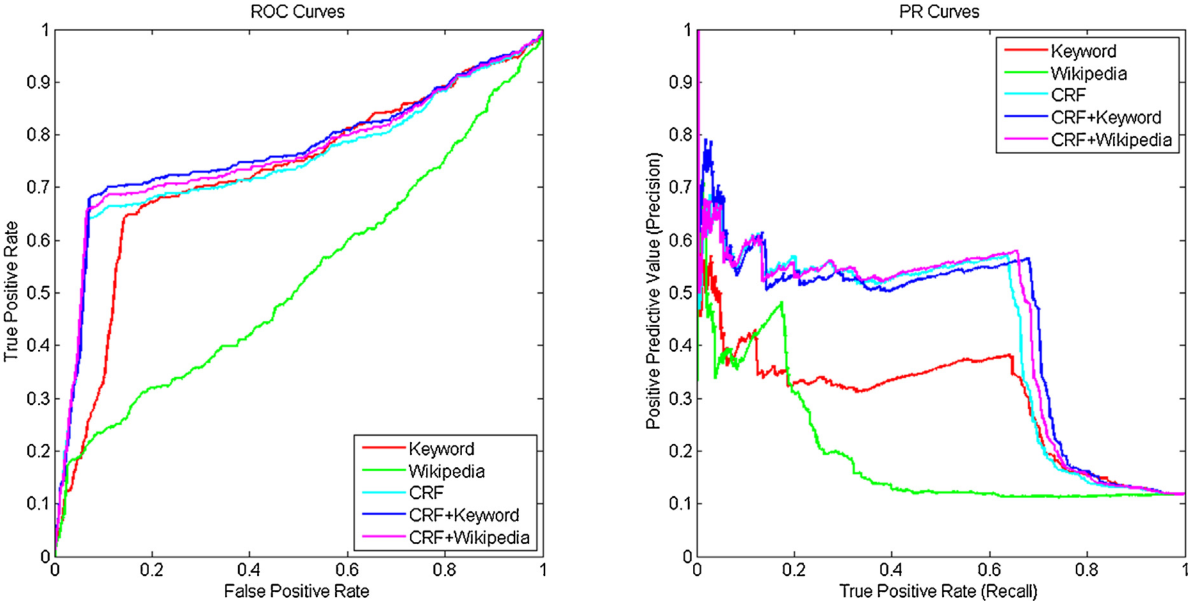


Fig. 3. ROC curves (left) and precision-recall curves (right) for the five methods.

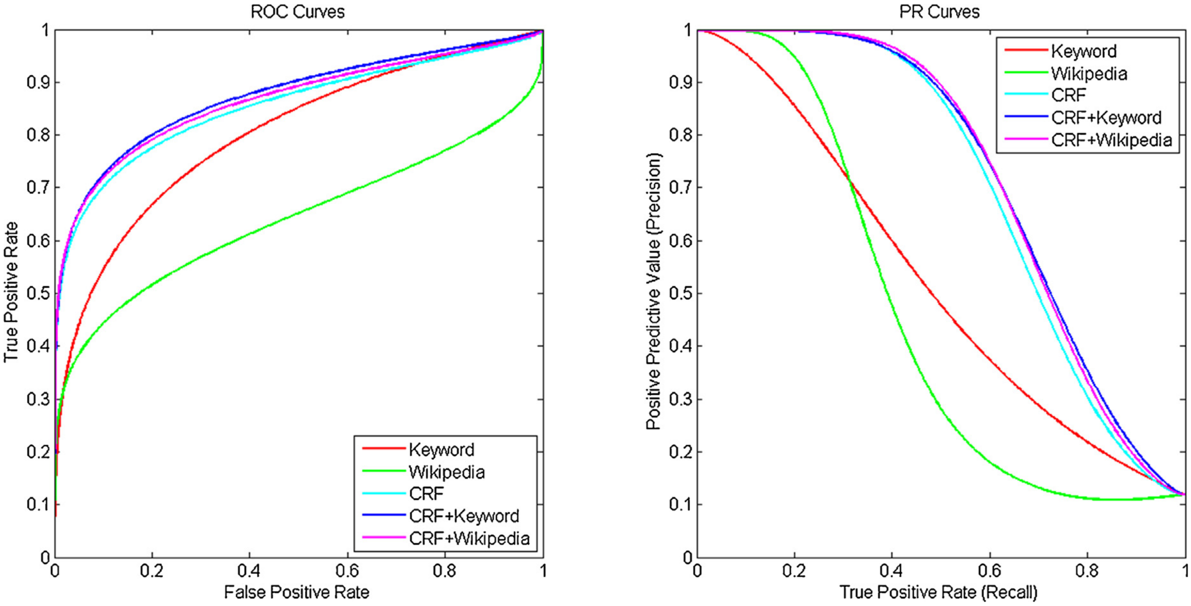


Fig. 4. Smoothed ROC curves (left) and precision-recall curves (right) for the five methods.

Wikipedia-based dictionary. By doing so, this paper has made the following contributions. First, it has applied the CRF-based statistical learning to the textual genre of scientific publications in computer science. This has enriched related entity extraction efforts that focused on analyzing scientific corpora (e.g., Prokofyev et al., 2014) as well as non-scientific corpora such as newspaper articles and email correspondence (e.g., Tang et al., 2006; Nadeau & Sekine, 2007; Sutton & McCallum, 2006). Second, bibliometrically, this paper has extended the unit of analysis to the entity-level from the classic author-, paper, and journal-level analyses (e.g., Yan & Ding, 2012). Conducting entity-level research has the advantage of gaining more granular insights on how various research-relevant entities relate to one another; it also helps refine investigations of transformative innovation, knowledge flow, and impact assessment. Finally, this paper has evaluated the performance of five approaches on entity extraction using appropriate indicators. The evaluative results will inform ongoing studies on content-aware informetric studies.

To construct domain specific vocabulary dictionaries, we have adopted two novel approaches—one included terms from papers' keywords and the other included Wikipedia entries in the category of computer science. As for the statistical models, we have employed CRF, a state-of-the-art design in entity extraction. Two features have also been integrated into the model, a keyword-based dictionary and a Wikipedia-based dictionary, both through the gazette function. The inclusion of vocabulary- and model-based approaches has enriched the word-based analyses in informetrics, such as co-word analysis (e.g., Callon, Courtial, & Laville, 1991; Milojević, Sugimoto, Yan, & Ding, 2011; Yan, Ding, & Jacob, 2012) and topic models (e.g., Blei, Ng, & Jordan, 2003; Blei & Lafferty, 2007; Ramage, Hall, Nallapati, & Manning, 2009), bringing a set of new ways to explore the content of scientific publications.

The advantage of the vocabulary-based approaches is that terms are readily available and requires fewer efforts to construct dictionaries. Nonetheless, the limitation is that the vocabulary-based approaches lack the ability to recognize entities that are not included in dictionaries. Meanwhile, keyword dictionaries are not controlled vocabularies—word- and dictionary-level congruence may be sparse. Statistical models, on the other hand, are able to discover new entities based on the conditional models, decision trees, hidden Markov models, bootstrapping, and other supervised or semi-supervised learning methods. Yet, as a supervised model, a training set is needed. A large, well annotated training set may take a considerably time to develop. Additionally, these training sets typically can only be applied to a specified domain. For instance, we cannot use the newspaper training set to train the CRF model for publications because they contain diverse bodies of entities. Similarly, the training set on computer science publications that developed in this study cannot be directly used to train the CRF model on other science domains. This disadvantage, therefore, may limit its applicability.

Through the use of several established evaluative indicators, this study has found that the model-based methods have outperformed the vocabulary-based ones, among which CRF with keyword-based dictionary has the best performance. Between the two vocabulary-based methods, the keyword-based one has a higher recall and the Wikipedia-based one has a higher precision. The result may be attributed to the fact that the constructed Wikipedia dictionary tended to contain a small set of well-defined terms but lacked a broad coverage. This has raised a challenge to construct a domain-specific dictionary that also possesses an extensive coverage on different subdomains. We see this as a future research direction.

Assessing different entity extraction approaches is the first step of our explorations toward a content-aware analysis of knowledge. The next step will be constructing knowledge networks using the identified entities to examine the provenance, trajectory, and popularity of various concepts, theories, and methods. These efforts will help provide empirical evidence to facilitate interdisciplinary, inter-organizational, and inter-territorial knowledge production and dissemination.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2015.04.003.

References

Akbik, A., & Löser, A. (2012). Kraken: N-ary facts in open information extraction. In *Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction* Association for Computational Linguistics, (pp. 52–56).

Alfonseca, E., & Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the first international conference on general WordNet* Mysore, India, (pp. 34–43).

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, *12*(4), 357–368.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction for the web. In *International joint conferences on artificial intelligence* (pp. 2670–2676), 7

Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, *3*(2).

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: A high-performance learning name-finder. In *Proceedings of the fifth conference on applied natural language processing* Association for Computational Linguistics, (pp. 194–201).

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1033.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the sixth workshop on very large corpora* , vol. 182.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The binormal assumption on precision-recall curves. In *20th international conference on pattern recognition (ICPR)* IEEE, (pp. 4263–4266).

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, *22*(1), 155–205.

Chagoyen, M., Carmona-Saez, P., Shatkay, H., Carazo, J. M., & Pascual-Montano, A. (2006). Discovering semantic features in the literature: A foundation for building functional associations. *BMC Bioinformatics*, *7*(41).

Christensen, J., Soderland, S., & Etzioni, O. (2010). Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 first international workshop on formalisms and methodology for learning by reading* Association for Computational Linguistics, (pp. 52–60).

Collier, N., Nobata, C., & Tsujii, J. I. (2000). Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th conference on computational linguistics* Association for Computational Linguistics, (pp. 201–207).

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora* (pp. 100–110).

DARPA. (2005). *Department of Defense fiscal year (FY) 2005 budget estimates*. DARPA. Retrieved Feb 13, 2015 from ⟨http://www.darpa.mil/WorkArea/DownloadAsset.aspx?id=1634⟩.

de Souza, C. R., Pizzolato, E. B., & dos Santos Anjo, M. (2012). Fingerspelling recognition with support vector machines and hidden conditional random fields. In *Advances in artificial intelligence—IBERAMIA 2012*. Berlin Heidelberg: Springer.

Del Corro, L., & Gemulla, R. (2013). ClausIE: Clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 355–366).

Demarest, B., & Sugimoto, C.R. (in press). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.23271.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817–842.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., et al. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, *165*(1), 91–134.

Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., et al. (2008). CoPub: A literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Research*, *36*(suppl 2), W406–W410.

Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J., & Alkema, W. (2010). Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Computational Biology*, *6*(9), e1000943.

Galvez, C., & de Moya-Anegón, F. (2012). A dictionary-based approach to normalizing gene names in one domain of knowledge from the biomedical literature. *Journal of Documentation*, *68*(1), 5–30.

Gupta, S., & Manning, C. D. (2014a). Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the 18th conference on computational natural language learning (CoNLL)*.

Gupta, S., & Manning, C. D. (2014b). SPIED: Stanford pattern-based information extraction and diagnostics. In *Proceedings of the ACL 2014 workshop on interactive language learning visualization, and interfaces (ACL-ILLVI)*.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

He, Y., & Kayaalp, M. (2008). Biological entity recognition with conditional random fields. In *AMIA annual symposium proceedings* American Medical Informatics Association, (pp. 293–297), vol. 2008.

Hirst, G. (1978). Discipline impact factors: Method for determining core journal lists. *Journal of the American Society for Information Science*, *29*(4), 171–172.

Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The unified medical language system: An informatics research collaboration. *Journal of the American Medical Informatics Association*, *5*(1), 1–11.

Hyland, K. (2000). *Disciplinary discourses: Social interaction in academic writing*. London: Longman.

Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: From information retrieval to biological discovery. *Nature Reviews Genetics*, *7*(2), 119–129.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., et al. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, *18*(5), 601–606.

Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications* Association for Computational Linguistics, (pp. 70–75).

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, *37*(6), 512–526.

Kripke, S. A. (1972). *Naming and necessity*. The Netherlands: Springer.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Liu, H., Hu, Z. Z., Torii, M., Wu, C., & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. *Journal of the American Medical Informatics Association*, *13*(5), 497–507.

Lou, W., & Qiu, J. (2014). Semantic information retrieval research based on co-occurrence analysis. *Online Information Review*, *38*(1), 4–23.

Lowe, H. J., & Barnett, G. O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association*, *271*(14), 1103–1108.

McCallum, A. (2002). Efficiently inducing features of conditional random fields. In *Proceedings of the nineteenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology*, *62*(10), 1933–1953.

Morgan, A. A., Hirschman, L., Colosimo, M., Yeh, A. S., & Colombe, J. B. (2004). Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, *37*(6), 396–410.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Nadeau, D., Turney, P., & Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Lecture Notes in Computer Science*, *4013*, 266–277.

Nallapati, R., & Manning, C. D. (2008). Legal docket-entry classification: Where machine learning stumbles. In *Proceedings of the conference on empirical methods in natural language processing* Association for Computational Linguistics, (pp. 438–446).

Nelson, S. J., Johnston, W. D., & Humphreys, B. L. (2001). Relationships in medical subject headings (MeSH). In *Relationships in the organization of knowledge*. The Netherlands: Springer.

Pawar, S., Srivastava, R., & Palshikar, G. K. (2012). Automatic gazette creation for named entity recognition and application to resume processing. In *Proceedings of the fifth ACM COMPUTE conference: Intelligent & scalable system technologies*. New York, NY: ACM Press.

Prokofyev, R., Demartini, G., & Cudré-Mauroux, P. (2014). Effective named entity recognition for idiosyncratic web collections. In *Proceedings of the 23rd international conference on World Wide Web* (pp. 397–408).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing* Association for Computational Linguistics, (pp. 248–256).

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the sixteenth national conference on artificial intelligence (AAAI-99)*. Menlo Park, CA: The AAAI Press.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on natural language learning at HLT-NAACL* Association for Computational Linguistics, (pp. 25–32).

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *International conference on language resources and evaluation* (pp. 1977–1980).

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In L. Getoor, & B. Taskar (Eds.), *Introduction to statistical relational learning* (pp. 93–128). Boston, MA: MIT Press.

Sutton, C., & McCallum, A. (2011). An introduction to conditional random fields. *Machine Learning*, *4*(4), 267–373.

Swanson, D. R. (1986). Fish oil Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18.

Swanson, D. R., Smalheiser, N. R., & Torvik, V. I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology*, *57*(11), 1427–1439.

Takeuchi, K., & Collier, N. (2005). Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, *33*(2), 125–137.

Tang, J., Hong, M., Li, J., & Liang, B. (2006). Tree-structured conditional random fields for semantic annotation. In *The semantic web—ISWC 2006*. Berlin Heidelberg: Springer.

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing* Association for Computational Linguistics, (pp. 214–221).

Thielen, C. (1995). *An approach to proper name tagging for German*. , arXiv preprint cmp-lg/9506024.

Torii, M., Hu, Z., Wu, C. H., & Liu, H. (2009). BioTagger-GM: A gene/protein name recognition system. *Journal of the American Medical Informatics Association*, *16*(2), 247–255.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks relate to each other. *Journal of the American Society for Information Science & Technology*, *63*(7), 1313–1326.

Yan, E., Ding, Y., & Jacob, E. K. (2012). Overlaying communities and topics: An analysis on publication networks. *Scientometrics*, *90*(2), 499–513.