Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach

Cornelia Caragea[1], **Florin Bulgarov**[1], **Andreea Godea**[1], **Sujatha Das Gollapalli**[2]
[1]Computer Science and Engineering, University of North Texas, TX, USA
[2]Institute for Infocomm Research, A*STAR, Singapore
ccaragea@unt.edu, FlorinBulgarov@my.unt.edu,
AndreeaGodea@my.unt.edu, gsdas@cse.psu.edu

Abstract

Given the large amounts of online textual documents available these days, e.g., news articles, weblogs, and scientific papers, *effective* methods for extracting keyphrases, which provide a high-level topic description of a document, are greatly needed. In this paper, we propose a supervised model for keyphrase extraction from research papers, which are embedded in citation networks. To this end, we design novel features based on citation network information and use them in conjunction with traditional features for keyphrase extraction to obtain remarkable improvements in performance over strong baselines.

1 Introduction

Keyphrase extraction is the problem of automatically extracting important phrases or concepts (i.e., the *essence*) of a document. Keyphrases provide a high-level topic description of a document and are shown to be rich sources of information for many applications such as document classification, clustering, recommendation, indexing, searching, and summarization (Jones and Staveley, 1999; Zha, 2002; Hammouda et al., 2005; Pudota et al., 2010; Turney, 2003). Despite the fact that keyphrase extraction has been widely researched in the natural language processing community, its performance is still far from being satisfactory (Hasan and Ng, 2014).

Many previous approaches to keyphrase extraction generally used only the textual content of a target document to extract keyphrases (Hulth, 2003; Mihalcea and Tarau, 2004; Liu et al., 2010). Recently, Wan and Xiao (2008) proposed a model that incorporates a local neighborhood of a document. However, their neighborhood is limited to textually-similar documents, where the cosine similarity between the *tf-idf* vectors of documents is used to compute their similarity. We posit that, in addition to a document's textual content and textually-similar neighbors, other informative neighborhoods exist that have the potential to improve keyphrase extraction. For example, in a scholarly domain, research papers are not isolated. Rather, they are highly inter-connected in giant *citation networks*, in which papers *cite* or *are cited* by other papers. In a citation network, information flows from one paper to another via the citation relation (Shi et al., 2010). This information flow and the influence of one paper on another are specifically captured by means of *citation contexts*, i.e., short text segments surrounding a citation's mention. These contexts are not arbitrary, but they serve as brief summaries of a cited paper. Figure 1 illustrates this idea using a small citation network of a paper by Rendle et al. (2010) that cites (Zimdars et al., 2001), (Hu et al., 2008), (Pan and Scholz, 2009) and (Shani et al., 2005) and is cited by (Cheng et al., 2013). The citation mentions and citation contexts are shown with a dashed line. Note the high overlap between the words in contexts and those in the title and abstract (shown in bold) and the author-annotated keywords.

One question that can be raised is the following: *Can we effectively exploit information available in large inter-linked document networks in order to improve the performance of keyphrase extraction?* The research that we describe in this paper addresses specifically this question using *citation networks of research papers* as a case study. Extracting keyphrases that can accurately "represent" research papers is crucial to dealing with the large numbers of research papers published during these "big data" times. The importance of keyphrase extraction from research papers is also emphasized by the recent SemEval 2010 Shared Task on this topic (Kim et al., 2010; Kim et al., 2013).

Our contributions. We present a supervised

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1435–1446,
October 25-29, 2014, Doha, Qatar.

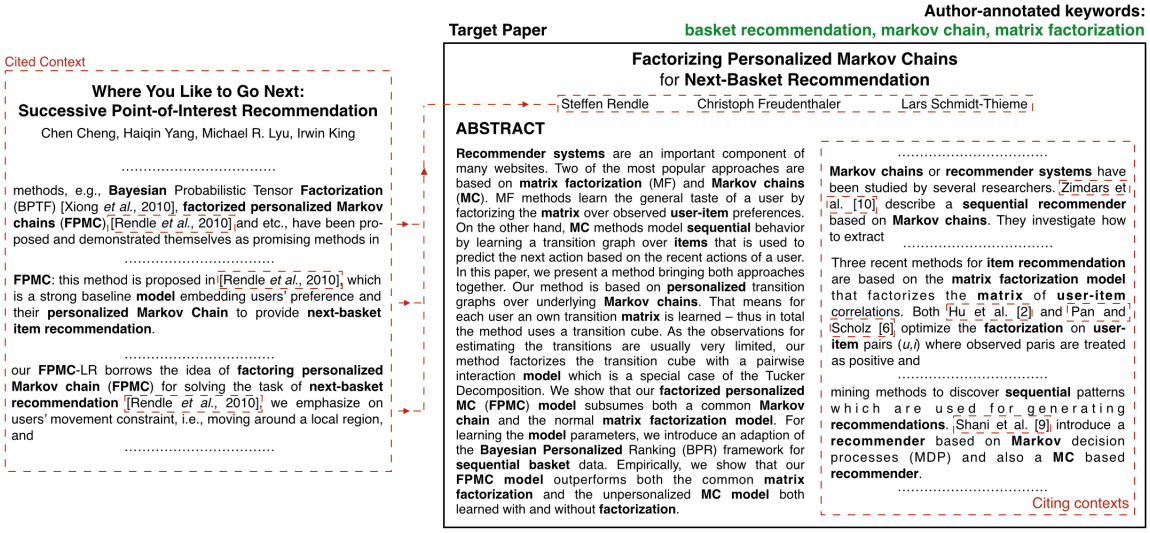

Figure 1: A small citation network corresponding to a paper by Rendle et al. (2010).

approach to keyphrase extraction from research papers that, in addition to the information contained in a paper itself, effectively incorporates, in the learned models, information from the paper's local neighborhood available in citation networks. To this end, we design novel features for keyphrase extraction based on citation context information and use them in conjunction with traditional features in a supervised probabilistic framework. We show empirically that the proposed models substantially outperform strong baselines on two datasets of research papers compiled from two machine learning conferences: the World Wide Web and Knowledge Discovery from Data.

The rest of the paper is organized as follows: We summarize closely related work in Section 2. The supervised classification for keyphrase extraction is discussed in Section 3. Experiments and results are presented in Section 4, followed by conclusions and future directions of our work.

2 Related Work

Many approaches to keyphrase extraction have been proposed in the literature along two lines of research: supervised and unsupervised, using different types of documents including scientific abstracts, newswire documents, meeting transcripts, and webpages (Frank et al., 1999; Hulth, 2003; Nguyen and Kan, 2007; Liu et al., 2009; Marujo et al., 2013; Mihalcea and Tarau, 2004).

In the supervised line of research, keyphrase extraction is formulated as a binary classification problem, where candidate phrases are classified as either positive (i.e., keyphrases) or negative (i.e., non-keyphrases) (Frank et al., 1999; Turney, 2000; Hulth, 2003). Different feature sets and classification algorithms gave rise to different models. For example, Hulth (2003) used four different features in conjunction with a *bagging* technique. These features are: term frequency, collection frequency, the relative position of the first occurrence and the part-of-speech tag of a term. Frank et al. (1999) developed a system called KEA that used only two features: *tf-idf* (term frequency-inverse document frequency) of a phrase and the *distance* of a phrase from the beginning of a document (i.e., its relative position) and used them as input to Naïve Bayes. Nguyen and Kan (2007) extended KEA to include features such as the distribution of keyphrases among different sections of a research paper, and the acronym status of a term. In contrast to these works, we propose novel features extracted from the local neighborhoods of documents available in interlinked document networks. Medelyan et al. (2009) extended KEA as well to integrate information from Wikipedia. In contrast, we used only information intrinsic to our data. Enhancing our models with Wikipedia information would be an interesting future direction to pursue.

In the unsupervised line of research, keyphrase extraction is formulated as a ranking problem, where keyphrases are ranked using their *tf* (Barker and Cornacchia, 2000), *tf-idf* (Zhang et al., 2007; Lee and Kim, 2008; Liu et al., 2009; Tonella et al., 2003), and *term informativeness* (Wu and Giles, 2013; Rennie and Jaakkola, 2005; Kireyev, 2009) (among others). The ranking based on *tf-idf* has

been shown to work well in practice (Liu et al., 2009; Hasan and Ng, 2010) despite its simplicity. Frantzi et al. (1998) combined linguistics and statistical information to extract technical terms from documents in digital libraries. Graph-based algorithms and centrality measures are also widely used in unsupervised models. A word graph is built for each document such that nodes correspond to words and edges correspond to word association patterns. Nodes are then ranked using graph centrality measures such as PageRank and its variants (Mihalcea and Tarau, 2004; Wan and Xiao, 2008; Liu et al., 2010; Zhao et al., 2011), HITS scores (Litvak and Last, 2008), as well as node degree and betweenness (Boudin, 2013; Xie, 2005). Wan and Xiao (2008) were the first to consider modeling a local neighborhood of a target document in addition to the document itself, and applied this approach to news articles on the Web. Their local neighborhood consists of textually similar documents, and did not capture information contained in document networks.

Using terms from citation contexts of scientific papers is not a new idea. It was used before in various applications. For example, Ritchie et al. (2006) used a combination of terms from citation contexts and existing index terms of a paper to improve indexing of cited papers. Citation contexts were also used to improve the performance of citation recommendation systems (Kataria et al., 2010; He et al., 2010) and to study author influence (Kataria et al., 2011). This idea of using terms from citation contexts resembles the analysis of hyperlinks and the graph structure of the Web, which are instrumental in Web search (Manning et al., 2008). Many current Web search engines build on the intuition that the anchor text pointing to a page is a good descriptor of its content, and thus use anchor text terms as additional index terms for a target webpage. The use of links and anchor text was thoroughly researched for IR tasks (Koolen and Kamps, 2010), broadening a user's search (Chakrabarti et al., 1998), query refinement (Kraft and Zien, 2004), and enriching document representations (Metzler et al., 2009).

Moreover, citation contexts were used for scientific paper summarization (Abu-Jbara and Radev, 2011; Qazvinian et al., 2010; Qazvinian and Radev, 2008; Mei and Zhai, 2008; Lehnert et al., 1990; Nakov et al., 2004). Among these, probably the most similar to our work is the work by Qazvinian et al. (2010), where a set of important keyphrases is extracted first from the citation contexts in which the paper to be summarized is cited by other papers and then the "best" subset of sentences that contain such keyphrases is returned as the summary. However, keyphrases in (Qazvinian et al., 2010) are extracted using frequent n-grams in a language model framework, whereas in our work, we propose a supervised approach to a different task: keyphrase extraction. Mei and Zhai (2008) used information from citation contexts to determine what sentences of a paper are of high impact (as measured by the influence of a target paper on further studies of similar or related topics). These sentences constitute the impact-based summary of the paper.

Despite the use of citation contexts and anchor text in many IR and NLP tasks, to our knowledge, we are the first to propose the incorporation of information available in citation networks for keyphrase extraction. In our recent work (Gollapalli and Caragea, 2014), we designed a fully unsupervised graph-based algorithm that incorporates evidence from multiple sources (citation contexts as well as document content) in a flexible manner to score keywords. In the current work, we present a supervised approach to keyphrase extraction from research papers that are embedded in large citation networks, and propose novel features that show improvement over strong supervised and unsupervised baselines. To our knowledge, features extracted from citation contexts have not been used before for keyphrase extraction in a supervised learning framework.

3 Problem Characterization

In citation networks, in addition to the information contained in a paper itself, *citing* and *cited* papers capture different aspects (e.g., topicality, domain of study, algorithms used) about the target paper (Teufel et al., 2006), with *citation contexts* playing an instrumental role. A citation context is defined as a window of n words surrounding a citation mention. We conjecture that citation contexts, which act as brief summaries about a cited paper, provide additional clues in extracting keyphrases for a target paper. These clues give rise to the unique design of our model, called citation-enhanced keyphrase extraction (CeKE).

3.1 Citation-enhanced Keyphrase Extraction

Our proposed citation-enhanced keyphrase extraction (CeKE) model is a supervised binary classifi-

Feature Name	Description
Existing features for keyphrase extraction	
tf-idf	term frequency * inverse document frequency, computed from a target paper; used in KEA
relativePos	the position of the first occurrence of a phrase divided by the total number of tokens; used in KEA and Hulth's methods
POS	the part-of-speech tag of the phrase; used in Hulth's methods
Novel features - Citation Network Based	
inCited	if the phrase occurs in cited contexts
inCiting	if the phrase occurs in citing contexts
citation tf-idf	the *tf-idf* value of the phrase, computed from the aggregated citation contexts
Novel features - Extensions of Existing Features	
first position	the distance of the first occurrence of a phrase from the beginning of a paper
tf-idf-Over	*tf-idf* larger than a threshold θ
firstPosUnder	the distance of the first occurrence of a phrase from the beginning of a paper is below some value β

Table 1: The list of features used in our model.

cation model, built on a combination of *novel* features that capture information from citation contexts and existing features from previous works. The features are described in §3.1.1. CeKE classifies candidate phrases as keyphrases (i.e., positive) or non-keyphrases (i.e., negative) using Naïve Bayes classifiers. Positive examples for training correspond to manually annotated keyphrases from the training research papers, whereas negative examples correspond to the remaining candidate phrases from these papers. The generation of candidate phrases is explained in §3.2.

Note that Naïve Bayes classifies a phrase as a keyphrase if the probability of the phrase belonging to the positive class is greater than 0.5. However, the default threshold of 0.5 can be varied to allow only high-confidence (e.g., 0.9 confidence) phrases to be classified as keyphrases.

3.1.1 Features

We consider the following features in our model, which are shown in Table 1. They are divided into three categories: (1) *Existing features for keyphrase extraction* include: *tf-idf*, i.e., the term frequency - inverse document frequency of a candidate phrase, computed for each target paper; This feature was used in KEA (Frank et al., 1999); *relative position*, i.e., the position of the first occurrence of a phrase normalized by the length (in the number of tokens) of the target paper; *POS*, i.e., a phrase's part-of-speech tag. If a phrase is composed by more than one term, then the POS will contain the tags of all terms. The relative position was used in both KEA and Hulth (2003), and POS was used in Hulth; (2) *Novel features - Citation Network Based* include: *inCited* and *inCiting*, i.e., boolean features that are true if the candidate phrase occurs in cited and citing contexts, respectively. We differentiate between cited and citing contexts for a paper: let d be a target paper and $\mathcal{C}$ be a citation network such that $d \in \mathcal{C}$. A cited context for d is a context in which d is cited by some paper d_i in $\mathcal{C}$. A citing context for d is a context in which d is citing some paper d_j in $\mathcal{C}$. If a paper is cited in multiple contexts by another paper, the contexts are aggregated into a single one; *citation tf-idf*, i.e., the tf-idf score of each phrase computed from the citation contexts; (3) *Novel features - Extend Other Existing Features* include: *first position* of a candidate phrase, i.e., the distance of the first occurrence of a phrase from the beginning of a paper; this is similar to relative position except that it does not consider the length of a paper; *tf-idf-Over*, i.e., a boolean feature, which is true if the *tf-idf* of a candidate phrase is greater than a threshold θ, and *firstPosUnder*, also a boolean feature, which is true if the distance of the first occurrence of a phrase from the beginning of a target paper is below some value β. This feature is similar to the feature *is-in-title*, used previously in the literature (Litvak and Last, 2008; Jiang et al., 2009). Both *tf-idf* and *citation tf-idf* features showed better results when each *tf* was divided by the maximum *tf* values from the target paper or citation contexts.

The *tf-idf* features have high values for phrases that are frequent in a paper or citation contexts, but are less frequent in collection and have low values for phrases with high collection frequency. We computed the *idf* component from each collection used in experiments. Phrases that occur in cited and citing contexts as well as early in a paper are likely to be keyphrases since: (1) they capture some aspect about the target paper and (2) authors start to describe their problem upfront.

3.2 Generating Candidate Phrases

We generate candidate phrases from the textual content of a target paper by applying parts-of-

Dataset	Num. (#) Papers	Average Cited Ctx.	Average Citing Ctx.	Average Keyphrases	#uni-grams	#bi-grams	#tri-grams
WWW	425	15.45	18.78	4.87	680	1036	247
KDD	365	12.69	19.74	4.03	363	853	189

Table 2: A summary of our datasets.

speech filters. Consistent with previous works (Hulth, 2003; Mihalcea and Tarau, 2004; Liu et al., 2010; Wan and Xiao, 2008), only nouns and adjectives are retained to form candidate phrases. The generation process consists of two steps. First, using the NLP Stanford part of speech tagger, we preprocess each document and keep only the nouns and adjectives corresponding to $\{NN, NNS, NNP, NNPS, JJ\}$. We apply the Porter stemmer on every word. The position of each word is kept consistent with the initial state of the document before any word removal is made.

Second, words extracted in the first step that have contiguous positions in a document are concatenated into n-grams. We used unigrams, bigrams, and trigrams ($n = 1, 2, 3$) as candidate phrases for classification. Similar to Wan and Xiao (2008), we eliminated phrases that end with an adjective and the unigrams that are adjectives.

4 Experiments and Results

In this section, we first describe our datasets and then present experimental design and results.

4.1 Datasets

In order to test the performance of our proposed approach, we built our own datasets since *citation-enhanced* evaluation benchmarks are not available for keyphrase extraction tasks. In particular, we compiled two datasets consisting of research papers from two top-tier machine learning conferences: World Wide Web (WWW) and Knowledge Discovery and Data Mining (KDD). Our choice for WWW and KDD was motivated by the availability of *author-input* keywords for each paper, which we used as gold-standard for evaluation.

Using the CiteSeer^x digital library[1], we retrieved the papers published in WWW and KDD (available in CiteSeer^x), and their citation network information, i.e., their cited and citing contexts. Since our goal is to study the impact of citation network information on extracting keyphrases, a paper was considered for analysis if it had at least one cited and one citing context. For each paper, we used: the title and abstract (referred to as the target paper) and its citation contexts. The reason for not considering the entire text of a paper is that scientific papers contain details, e.g., discussion of results, experimental design, notation, that do not provide additional benefits for extracting keyphrases. Hence, similar to (Hulth, 2003; Mihalcea and Tarau, 2004; Liu et al., 2009), we did not use the entire text of a paper. However, extracting keyphrases from sections such as "introduction" or "conclusion" needs further attention.

From the pdf of each paper, we extracted the author-input keyphrases. An analysis of these keyphrases revealed that generally authors describe their work using, almost half of the time, bigrams, followed by unigrams and only rarely using trigrams (or higher n-grams). A summary of our datasets that contains the number of papers, the average number of cited and citing contexts per paper, the average number of keyphrases per paper, and the number of unigrams, bigrams and trigrams, in each collection, is shown in Table 2.

Consistent with previous works (Frank et al., 1999; Hulth, 2003), the positive and negative examples in our datasets correspond to candidate phrases that consist of up to three tokens. The positive examples are candidate phrases that have a match in the author-input keyphrases, whereas negative examples correspond to the remaining candidate phrases.

Context lengths. In CiteSeer^x, citation contexts have about 50 words on each side of a citation mention. A previous study by Ritchie et al. (2008) shows that a fixed window length of about 100 words around a citation mention is generally effective for information retrieval tasks. For this reason, we used the contexts provided by CiteSeer^x directly. However, in future, it would be interesting to incorporate in our models more sophisticated approaches to identifying the text that is relevant to a target citation (Abu-Jbara and Radev, 2012; Teufel, 1999) and study the influence of context lengths on the quality of extracted keyphrase.

[1] http://citeseerx.ist.psu.edu/

	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
Citation - Enhanced (CeKE)	**0.227**	**0.386**	**0.284**	**0.213**	**0.413**	**0.280**
Hulth - n-gram with tags	0.165	0.107	0.129	0.206	0.151	0.172
KEA	0.210	0.146	0.168	0.178	0.124	0.145

Table 3: The comparison of CeKE with supervised approaches on WWW and KDD collections.

4.2 Experimental Design

Our experiments are designed around the following research questions:

1. *How does the performance of citation-enhanced keyphrase extraction (CeKE) compare with the performance of existing supervised models that use only information intrinsic to the data and what are the most informative features for classification?* We compared CeKE's performance with that of classifiers trained on KEA features only and Hulth's features only and present a ranking of features based on information gain.
2. *How do supervised models that integrate citation network information compare with recent unsupervised models?* Since recent unsupervised approaches are becoming competitive with supervised approaches (Hasan and Ng, 2014), we also compared CeKE with unsupervised ranking of candidate phrases by TF-IDF, TextRank (Mihalcea and Tarau, 2004) and ExpandRank (Wan and Xiao, 2008). For unsupervised, we considered top 5 and top 10 ranked phrases when computing "@5" and "@10" measures.
3. *How well does our proposed model perform in the absence of either cited or citing contexts?* Since newly published scientific papers are not cited by many other papers, e.g., due to their recency, no cited contexts are available. We studied the quality of predicted keyphrases when either cited or citing contexts are missing. For this, we compared the performance of models trained using both cited and citing contexts with that of models that use either cited or citing contexts.

Evaluation metrics. To evaluate the performance of CeKE, we used the following metrics: precision, recall and F1-score for the positive class since correct identification of keyphrases is of most interest. These metrics were widely used in previous works (Hulth, 2003; Mihalcea and Tarau, 2004; Wan and Xiao, 2008; Hasan and Ng, 2010). The reported values are averaged in 10-fold cross-validation experiments, where folds were created at document level and candidate phrases were extracted from the documents in each fold to form the training and test sets. In all experiments, we used Naïve Bayes and their Weka implementation[2]. However, any probabilistic classifier that returns a posterior probability of the class given an example, can be used with our features.

The θ parameter was set to the (title and abstract) *tf-idf averaged* over the entire collection, while β was set to 20. These values were estimated on a validation set sampled from training.

4.3 Results and Discussion

The impact of citation network information on the keyphrase extraction task. Table 3 shows the results of the comparison of CeKE with two supervised approaches, KEA and Hulth's approach. The features used in KEA are the *tf-idf* and the *relative position* of a candidate phrase, whereas those used in Hulth's approach are *tf*, *cf* (i.e., collection frequency), *relative position* and *POS tags*. CeKE is trained using all features from Table 1. Among the three methods for candidate phrase formation proposed in Hulth (2003), i.e., n-grams, NP-chunks, and POS Tag Patterns, our Hulth's implementation is based on n-grams since this gives the best results among all methods (see (Hulth, 2003) for more details). In addition, the n-grams method is the most similar to our candidate phrase generation and that used in Frank et al. (1999).

As can be seen from Table 3, CeKE outperforms KEA and Hulth's approach in terms of all performance measures on both WWW and KDD, with a substantial improvement in recall over both approaches. For example, on WWW, CeKE achieves a recall of 0.386 compared to 0.146 and 0.107 recall achieved by KEA and Hulth's, respectively.

[2]http://www.cs.waikato.ac.nz/ml/weka/

	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
Citation - Enhanced (CeKE)	**0.227**	**0.386**	**0.284**	**0.213**	**0.413**	**0.280**
TF-IDF - Top 5	0.089	0.100	0.094	0.083	0.102	0.092
TF-IDF - Top 10	0.075	0.169	0.104	0.080	0.203	0.115
TextRank - Top 5	0.058	0.071	0.062	0.051	0.065	0.056
TextRank - Top 10	0.062	0.133	0.081	0.053	0.127	0.072
ExpandRank - 1 neigh. - Top 5	0.088	0.109	0.095	0.077	0.103	0.086
ExpandRank - 1 neigh. - Top 10	0.078	0.165	0.101	0.071	0.177	0.098
ExpandRank - 5 neigh. - Top 5	0.093	0.113	0.100	0.080	0.108	0.090
ExpandRank - 5 neigh. - Top 10	0.080	0.172	0.104	0.068	0.172	0.095
ExpandRank - 10 neigh. - Top 5	0.094	0.113	0.100	0.077	0.103	0.086
ExpandRank - 10 neigh. - Top 10	0.076	0.162	0.099	0.065	0.164	0.091

Table 5: The comparison of CeKE with unsupervised approaches on WWW and KDD collections.

Rank	*Feature*	IG Score
1	*abstract tf-idf*	0.0234
2	*first position*	0.0188
3	*citation tf-idf*	0.0177
4	*relativePos*	0.0154
5	*firstPosUnder*	0.0148
6	*inCiting*	0.0129
7	*inCited*	0.0098
8	*POS*	0.0085
9	*tf-idf-Over*	0.0078

Table 4: Feature ranking by Info Gain on WWW.

Although there are only small variations from KEA to Hulth's approach, KEA performs better on WWW, but worse on KDD compared with Hulth's approach. In contrast, CeKE shows consistent improvement over the two approaches on both datasets, hence, effectively making use of the information available in the citation network.

In order to understand the importance of our features, we ranked them based on Information Gain (IG), which determines how informative a feature is with respect to the class variable. Table 4 shows the features ranked in decreasing order of their IG scores for WWW. As can be seen from the table, *tf-idf* and *citation tf-idf* are both highly ranked, first and third, respectively, illustrating that they contain significant information in predicting keyphrases. The *first position* of a phrase is also of great impact. This is consistent with the fact that almost half of the identified keywords and about 20% of the annotated keyphrases appear in title. Similar ranking is obtained on KDD.

The comparison of CeKE with unsupervised state-of-the-art models. Table 5 shows the results of the comparison of CeKE with three unsupervised ranking approaches: TF-IDF (Tonella et al., 2003), TextRank (Mihalcea and Tarau, 2004), and ExpandRank (Wan and Xiao, 2008). TF-IDF and TextRank use information only from the target paper, whereas ExpandRank uses a small textual neighborhood in addition to the target paper. Note that, for all unsupervised methods, we used Porter stemmer and the same candidate phrase generation as in CeKE, as explained in §3.2.

For TF-IDF, we first tokenized the target paper and computed the score for each word, and then formed phrases and summed up the score of every word within a phrase. For TextRank, we built an undirected graph for each paper, where the nodes correspond to words in the target paper and edges are drawn between two words that occur next to each other in the text, i.e., the window size is 2. For ExpandRank, we built an undirected graph for each paper and its local textual neighborhood. Again, nodes correspond to words in the target paper and its textually similar papers and edges are drawn between two words that occur within a window of 10 words from each other in the text, i.e., the window size is 10. We performed experiments with 1, 5, and 10 textually-similar neighbors. For TextRank and ExpandRank, we summed up the scores of words within a phrase as in TF-IDF.

	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
CeKE - Both contexts	**0.227**	**0.386**	**0.284**	**0.213**	**0.413**	**0.280**
CeKE - Only cited contexts	0.222	0.286	0.247	0.192	0.300	0.233
CeKE - Only citing contexts	0.203	0.342	0.253	0.195	0.351	0.250

Table 6: Results of CeKE using both contexts and using with only cited or citing contexts.

For each unsupervised method, we computed results for top 5 and top 10 ranked phrases. As can be seen from Table 5, CeKE substantially outperforms all the other methods for our domain of study, i.e., papers from WWW and KDD, illustrating again that the citation network of a paper contains important information that can show remarkable benefits for keyphrase extraction. Among all unsupervised methods, ExpandRank with fewer textual similar neighbor (one or five) performs the best. This is generally consistent with the results shown in (Wan and Xiao, 2008) for news articles.

The effect of cited and citing contexts information on models' performance. Table 6 shows the precision, recall and F-score values for some variations of our method when: (1) all the citation contexts for a paper are used, (2) only cited contexts are used, (3) only citing contexts are used. The motivation behind this experiment was to determine how well the proposed model would perform on newly published research papers that have not accumulated citations yet. As shown in the table, there is no substantial difference in terms of precision between CeKE models that use only cited or only citing contexts, although the recall is substantially higher for the case when only citing contexts are used, for both WWW and KDD. The CeKE that uses both citing and cited contexts achieves a substantially higher recall and only a slightly higher precision compared with the cases when only one context type is available. The fact that the citing context information provides a slight improvement in performance over cited contexts is consistent with the intuition that when citing a paper y, an author generally summarizes the main ideas from y using important words from a target paper x, making the citing contexts to have higher overlap with words from x. In turn, a paper z that cites x may use paraphrasing to summarize ideas from x with words more similar to those from z.

Note that the results of all above experiments are statistically significant at p-values ≥ 0.05, using a paired t-test on F1-scores.

4.4 Anecdotal Evidence

In order to check the transferability of our proposed approach to other research fields, e.g., natural language processing, it would be interesting to use our trained classifiers on WWW and KDD collections and evaluate them on new collections such as NLP related collections. Since NLP collections annotated with keyphrases are not available, we show anecdotal evidence for only one paper. We selected for this task an award winning paper published in the EMNLP conference. The paper's title is "Unsupervised semantic parsing" and has won the Best Paper Award in the year 2009 (Poon and Domingos, 2009). In order for our algorithm to work, we gathered from the Web (using Google Scholar) all the cited and citing contexts that were available (49 cited contexts and 30 citing contexts). We manually annotated the target paper with keyphrases. The title, abstract and all the contexts were POS tagged using the NLP Stanford tool. We then trained a classifier on the features shown in Table 1, on both WWW and KDD datasets combined. The trained classifier was used to make predictions, which were compared against the manually annotated keyphrases. The results are shown in Figure 2, which displays the title and abstract of the paper and the predicted keyphrases. Candidate phrases that are predicted as keyphrases are marked in red bold, those predicted as non-keyphrases are shown in black, while the filtered out words are shown in light gray.

We tuned our classifier trained on WWW and KDD to return as keyphrases only those that had an extremely high probability to be keyphrases. Specifically, we used a threshold of 0.985. The probability of each returned keyphrase (which is above 0.985) is shown in the upper right corner of a keyphrase. Human annotated keyphrases are marked in italic, under the figure. There is a clear match between the predictions and the human annotations. It is also possible to extract more or less keyphrases simply by adjusting the threshold

Unsupervised Semantic Parsing$^{0.997}$

We present the first unsupervised approach to the problem of learning a **semantic parser**$^{1.000}$, using **Markov logic**$^{0.991}$. Our **USP system**$^{0.985}$ transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic variations of the same meaning. The MAP **semantic parse**$^{1.000}$ of a sentence is obtained by recursively assigning its parts to lambda-form clusters and composing them. We evaluate our approach by using it to extract a knowledge base from biomedical abstracts and answer questions. **USP**$^{1.000}$ substantially outperforms TextRunner, DIRT and an informed baseline on both precision and recall on this task.

Human annotated labels: *unsupervised semantic parsing, Markov logic, USP system*

Figure 2: The title and abstract of an EMNLP paper by Poon and Domingos (2009) and human annotated keyphrases for the paper. Black words represent candidate phrases. Red bold words represent predicted keyphrases. The numbers above predicted keyphrases are probabilities for the positive class assignment.

on the probability output by Naïve Bayes. For example, if we decrease the threshold to 0.920 the following phrases would be added to the returned set of keyphrases: *dependency trees, quasi-logical forms* and *unsupervised approach*.

Another interesting aspect is the frequency of occurrence of the predicted keyphrases in the cited and citing contexts. Table 7 shows the term-frequency of every predicted keyphrase within the citation network. For example, the phrase *semantic parser* appears in 29 cited contexts and 26 citing contexts. The reason for the higher cited context frequency is not necessarily due to importance, but could be due to the larger number of cited vs. citing contexts for this paper (49 vs. 30). The high rate of keyphrases within the citation network validates our assumption of the importance of citation networks for keyphrase extraction.

Finally, we performed the same experiment with Hulth's and KEA methods. While the classifier trained on Hulth's features did not identify any keyphrases, KEA managed to identify several good ones (e.g., *USP*, *semantic parser*), but left out some important ones (e.g., *Markov logic, unsupervised*). Moreover, the keyphrases predicted by KEA have a lower confidence. For this reason, lowering the probability threshold would result in selecting other bad keyphrases.

4.5 Error analysis

We performed an error analysis and found that candidate phrases are predicted as keyphrases (FPs), although they do not appear in gold standard, i.e., the set of author-input keyphrases, in cases when: 1) a more general terms is used to describe an important concept of a document, e.g., *co-authorship prediction* represented as *link prediction* or *Twitter platform* represented as *social media*; 2) an important concept is omitted (either intentionally or forgetfully) from the set of author-input keyphrases.

Keyphrase	#cited c.	#citing c.
semantic parser	29	26
USP	31	10
Markov logic	15	10
unsupervised semantic parsing	12	1
USP system	3	2

Table 7: Frequency of the predicted keyphrases in cited / citing contexts.

Hence, while we believe that authors are the best keyphrase annotators for their own work, there are cases when important keyphrases are overlooked or expressed in different ways, possibly due to the human subjective nature in choosing important keyphrases that describe a document. To this end, a limitation of our model is the use of a single gold standard keyphrase annotation. In future, we plan to acquire several human keyphrase annotation sets for our datasets and test the performance of the proposed approach on these annotation sets, independently and in combination.

Keyphrases that appear in gold standard are predicted as non-keyphrases (FNs) when: 1) a keyphrase is infrequent in abstract; 2) its distance from the beginning of a document is large; 3) does not occur or occurs only rarely in a document's citation contexts, either citing or cited contexts. Examples of FNs are model/algorithm/approach names, e.g., *random walks*, that appear in sentences such as: "In this paper, we model the problem [· · ·] by using *random walks*." Although such

a sentence may appear further away from the beginning of an abstract, it contains significant information from the point of view of keyphrase extraction. The design of patters such as $< by\ using\ \$model >$ or $< uses\ \$model >$ could lead to improved classification performance.

Further investigation of FPs and FNs will be considered in future work. We believe that a better understanding of errors has the potential to advance *state-of-the-art* for keyphrase extraction.

5 Conclusion and Future Directions

In this paper, we presented a supervised classification model for keyphrase extraction from scientific research papers that are embedded in citation networks. More precisely, we designed novel features that take into account citation network information for building supervised models for the classification of candidate phrases as keyphrases or non-keyphrases. The results of our experiments show that the proposed supervised model trained on a combination of citation-based features and existing features for keyphrase extraction performs substantially better compared with state-of-the-art supervised and unsupervised models.

Although we illustrated the benefits of leveraging inter-linked document networks for keyphrase extraction from scientific documents, the proposed model can be extended to other types of documents such as webpages, emails, and weblogs. For example, the anchor text on hyperlinks in weblogs can be seen as the "citation context".

Another aspect of future work would be the use of external sources to better identify candidate phrases. For example, the use of Wikipedia was studied before to check if the concept behind a phrase has its own Wikipedia page (Medelyan et al., 2009). Furthermore, since citations occur in all sciences, extensions of the proposed model to other domains, e.g., Biology and Chemistry, and other applications, e.g., document summarization, similar to Mihalcea and Tarau (2004) and Qazvinian et al. (2010), are of particular interest.

Acknowledgments

We are grateful to Dr. C. Lee Giles for the CiteSeerX data, which allowed the generation of citation graphs. We also thank Kishore Neppalli and Juan Fernández-Ramízer for their help with various dataset construction tasks. We very much appreciate the constructive feedback from our anonymous reviewers. This research was supported in part by NSF awards #1353418 and #1423337 to Cornelia Caragea. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 500–509.

Amjad Abu-Jbara and Dragomir Radev. 2012. Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 80–90.

Ken Barker and Nadia Cornacchia. 2000. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, AI '00, pages 40–52, London, UK, UK. Springer-Verlag.

Florian Boudin. 2013. A comparison of centrality measures for graph-based keyphrase extraction. In *Proc. of IJCNLP*, pages 834–838, Nagoya, Japan.

Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. 1998. Automatic resource compilation by analyzing hyperlink structure and associated text. *Comput. Netw. ISDN Syst.*, 30(1-7):65–74, April.

Chen Cheng, Haiqin Yang, Michael R. Lyu, and Irwin King. 2013. Where you like to go next: Successive point-of-interest recommendation. In *Proc. of IJCAI'13*, pages 2605–2611, Beijing, China.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 668–673, Stockholm, Sweden.

Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *Proc. of ECDL '98*, pages 585–604.

Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting keyphrases from research papers using citation networks. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-14)*, Québec City, Québec, Canada.

Khaled M. Hammouda, Diego N. Matute, and Mohamed S. Kamel. 2005. Corephrase: Keyphrase extraction for document clustering. In *Proc. of the 4th*

International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM'05, pages 265–274, Leipzig, Germany.

Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 365–373.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proc. of WWW '10*, pages 421–430, Raleigh, North Carolina, USA.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proc. of the 8th IEEE Intl. Conference on Data Mining*, ICDM '08, pages 263–272.

Anette Hulth. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 216–223.

Xin Jiang, Yunhua Hu, and Hang Li. 2009. A Ranking Approach to Keyphrase Extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 756–757. ACM.

Steve Jones and Mark S. Staveley. 1999. Phrasier: A system for interactive document retrieval using keyphrases. In *Proceedings of SIGIR '99*, pages 160–167, Berkeley, California, USA.

Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing context in generative bayesian models for linked corpus. In *In Proc. of AAAI '10*, pages 1340–1345, Atlanta, Georgia, USA.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. 2011. Context sensitive topic models for author influence in document networks. In *Proceedings of IJCAI'11*, pages 2274–2280, Barcelona, Catalonia, Spain.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 21–26, Los Angeles, California.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2013. Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation, Springer*, 47(3):723–742.

Kirill Kireyev. 2009. Semantic-based estimation of term informativeness. In *Proc. of NAACL '09*, pages 530–538, Boulder, Colorado.

Marijn Koolen and Jaap Kamps. 2010. The importance of anchor text for ad hoc search revisited. In *Proceedings of SIGIR '10*, pages 122–129, Geneva, Switzerland.

Reiner Kraft and Jason Zien. 2004. Mining anchor text for query refinement. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 666–674, New York, NY, USA. ACM.

Sungjick Lee and Han-joon Kim. 2008. News Keyword Extraction for Topic Tracking. In *Proceedings of the 2008 Fourth International Conference on Networked Computing and Advanced Information Management - Volume 02*, NCM '08, pages 554–559, Washington, DC, USA. IEEE Computer Society.

Wendy Lehnert, Claire Cardie, and Ellen Rilofl. 1990. Analyzing research papers using citation sentences. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 511–518.

Marina Litvak and Mark Last. 2008. Graph-Based Keyword Extraction for Single-Document Summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pages 17–24, Manchester, United Kingdom.

Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009. Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts. In *Proceedings of NAACL '09*, pages 620–628, Boulder, Colorado.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic Keyphrase Extraction via Topic Decomposition. In *Proceedings of EMNLP '10*, pages 366–376, Cambridge, Massachusetts.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Luís Marujo, Ricardo Ribeiro, David Martins de Matos, João Paulo Neto, Anatole Gershman, and Jaime G. Carbonell. 2013. Key phrase extraction of lightly filtered broadcast news. *CoRR*.

Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1318–1327, Singapore.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio.

Donald Metzler, Jasmine Novak, Hang Cui, and Srihari Reddy. 2009. Building enriched document representations using aggregated anchor text. In *Proc. of SIGIR '09*, pages 219–226, Boston, MA, USA.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain.

Preslav I. Nakov, Ariel S. Schwartz, and Marti A. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *SIGIR Workshop on Search and Discovery in Bioinformatics*.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase Extraction in Scientific Publications. In *Proc. of the Intl. Conf. on Asian digital libraries*, ICADL'07, pages 317–326, Hanoi, Vietnam.

Rong Pan and Martin Scholz. 2009. Mind the gaps: Weighting the unknown in large-scale one-class collaborative filtering. In *Proceedings of KDD '09*, pages 667–676, Paris, France.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 1–10, Singapore.

Nirmala Pudota, Antonina Dattolo, Andrea Baruzzo, Felice Ferrara, and Carlo Tasso. 2010. Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems*.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proc. of the 22nd Intl. Conference on Computational Linguistics*, COLING '08, pages 689–696, Manchester, United Kingdom.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 895–903.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *WWW '10*, pages 811–820, Raleigh, North Carolina.

Jason D. M. Rennie and Tommi Jaakkola. 2005. Using Term Informativeness for Named Entity Detection. In *Proc. of SIGIR '05*, pages 353–360.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006. How to find better index terms through citations. In *Proc. of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, CLIIR '06, pages 25–32, Sydney, Australia.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proc. of CIKM '08*, pages 213–222, Napa Valley, California, USA.

Guy Shani, David Heckerman, and Ronen I. Brafman. 2005. An mdp-based recommender system. *J. Mach. Learn. Res.*, 6:1265–1295, December.

Xiaolin Shi, Jure Leskovec, and Daniel A. McFarland. 2010. Citing for high impact. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 49–58, Gold Coast, Queensland, Australia.

S. Teufel, A. Siddharthan, and D. Tidhar. 2006. Automatic classification of citation function. In *Proceedings of EMNLP-06*.

S. Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Paolo Tonella, Filippo Ricca, Emanuele Pianta, and Christian Girardi. 2003. Using Keyword Extraction for Web Site Clustering. In *Web Site Evolution, 2003. Theme: Architecture. Proceedings. Fifth IEEE International Workshop on*, pages 41–48.

Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2.

Peter D. Turney. 2003. Coherent Keyphrase Extraction via Web Mining. In *Proceedings of the 18th international joint conference on Artificial intelligence*, IJCAI'03, pages 434–439, Acapulco, Mexico.

Xiaojun Wan and Jianguo Xiao. 2008. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of AAAI '08*, pages 855–860, Chicago, Illinois.

Zhaohui Wu and Lee C. Giles. 2013. Measuring term informativeness in context. In *Proceedings of NAACL '13*, pages 259–269, Atlanta, Georgia.

Zhuli Xie. 2005. Centrality Measures in Text Mining: Prediction of Noun Phrases that Appear in Abstracts. In *Proceedings of the ACL Student Research Workshop*, pages 103–108, Ann Arbor, Michigan.

Hongyuan Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR*.

Yongzheng Zhang, Evangelos Milios, and Nur Zincir-Heywood. 2007. A Comparative Study on Key Phrase Extraction Methods in Automatic Web Site Summarization. *Journal of Digital Information Management*, 5(5):323.

Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. 2011. Topical Keyphrase Extraction from Twitter. In *Proceedings of HLT '11*, pages 379–388, Portland, Oregon.

Andrew Zimdars, David Maxwell Chickering, and Christopher Meek. 2001. Using temporal data for making recommendations. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 580–588.