Natural Language Engineering **19** (4): 481–515. © Cambridge University Press 2012
doi:10.1017/S1351324912000277

Context identification of sentences in research articles: Towards developing intelligent tools for the research community

M. A. ANGROSH, STEPHEN CRANEFIELD
and NIGEL STANGER

Department of Information Science, University of Otago, Dunedin, New Zealand
e-mail: {angrosh, scranefield, nstanger}@infoscience.otago.ac.nz

(*Received 23 September 2011; revised 30 August 2012; accepted 4 September 2012; first published online 10 October 2012*)

Abstract

Scientific literature is an important medium for disseminating scientific knowledge. However, in recent times, a dramatic increase in research output has resulted in challenges for the research community. An increasing need is felt for tools that exploit the full content of an article and provide insightful services with value beyond quantitative measures such as impact factors and citation counts. However, the intricacies of language and thought, and the unstructured format of research articles present challenges in providing such services. The identification of sentence contexts that encode the role of specific sentences in advancing an article's scientific argument can facilitate in developing intelligent tools for the research community. This paper describes our research work in this direction. First, we investigate the possibility of identifying contexts associated with sentences and propose a scheme of thirteen context type definitions for sentences, based on the generic rhetorical pattern found in scientific articles. We then present the results of our experiments using sequential classifiers – conditional random fields – for achieving automatic context identification. We also describe our Semantic Web application developed for providing citation context based information services for the research community. Finally, we present a comparison and analysis of our results with similar studies and explain the distinct features of our application.

1 Introduction

Over the years, scientific articles have been the primary medium for disseminating scientific knowledge. Typically, the published scientific information builds upon the previous or existing scientific knowledge and citations play a key role in defining these inter-article relationships. One of the requirements for conducting effective research is the identification of these inter-article relations, which helps in avoiding duplication and also to position one's research with respect to previous research. In recent times, a dramatic increase in research output (Gaillard 2008; Research4Life 2009) has resulted in difficulties in identifying these relations. For example, by using current information services, it is possible to identify the number of documents citing a given document; but it may not be possible to identify the reasons for

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

which the citing documents have cited the given document. As Buckingham Shum *et al.* point out, present day information systems fail to answer some of the basic questions of critical inquiry such as providing information about documents that either support or challenge a given document or facilitate tracing the intellectual lineage for a given idea (Buckingham Shum *et al.* 2007). An increasing need is felt for intelligent tools capable of providing insightful services that provide value beyond quantitative measures such as impact factors and citation counts. However, the intricacies of language and the unstructured format of research articles present challenges in providing such services. Identifying contexts associated with sentences can help in designing information tools for the research community that identify inter-article relationships. For example, it would be easier to identify the gaps in the works cited in an article.

Against this background, this research is taken up to investigate the possibility of identifying context types associated with sentences and using this information for providing citation context based information services for the research community. The task of designing citation context based information services involves a series of tasks. We describe in this paper the following steps followed in our study for designing such services.

(1) Defining context types for sentences in academic articles
(2) Carrying out supervised learning methods for achieving automatic context identification
(3) Creating structured information from unstructured text
(4) Developing an application that uses structured information for providing citation context based information services

The structure of this paper is as follows. In Section 2, we identify the related work in the field of scientific discourse annotation and present our rationale for defining a new set of context type definitions and the use of sequential classifiers. In Section 3, we present our scheme of thirteen context type definitions for sentences in scientific articles and explain the procedure adopted for deriving these definitions. We also compare the definitions with similar studies in this section. In Section 4, we briefly explain our information extraction system developed for text preparation tasks. We explain, in Section 5, our experiments carried out with conditional random fields (CRFs) for achieving automatic context identification. In Section 6, we describe the different citation context based information services provided by the Semantic Web application developed in our study. In Section 8, we present an analysis of the features used in machine-learning experiments and compare the classifier results with results of similar studies. We also discuss the distinct features provided by our application. Finally, we conclude this paper by providing pointers to future work.

2 Related work

Studies on sentence classification in scientific articles have focused on identifying the functions associated with sentences and the rhetorical relations between them. Some studies have considered the full text of an article and others have limited their

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

focus either to sentences with citations or abstracts of an article. We describe in the following sections different studies in these areas.

2.1 Annotation of full-text documents

Over the years, several studies have investigated the task of annotating scientific texts. Teufel (1999) focused on analyzing scientific text with what she called 'Argumentative Zoning' and proposed a classification scheme of seven categories (argumentative zones) for sentences in scientific articles, based on the argumentation involved. Mizuta and Collier (2004a) extended Teufel's (1999) scheme for carrying out zone analysis in biology texts. Mizuta and Collier (2004b) further provided theoretical and practical support for their zone analysis. Langer, Lüngen, and Bayerl (2004) noted that newer applications in areas such as the Semantic Web required richer and more fine-grained annotation of documents and presented a scheme involving sixteen topic types. Considering the scheme of Mizuta and Collier (2004a) as complex, Wilbur, Rzhetsky, and Shatkay (2006) devised an annotation scheme of five categories for biomedical texts focusing on the semantic content of the sentence and presented guidelines for annotation of text without using the semantic content of the sentence.

Shatkay *et al.* (2008) used the scheme of Wilbur *et al.* (2006) and conducted machine-learning experiments for multidimensional classification of biomedical texts. Liakata *et al.* (2010) presented and applied two complementary annotation schemes for scientific papers in Chemistry: the Core Scientific Concepts (CoreSC) annotation scheme based on the CISP (the proposed Core Information about Scientific Papers) metadata (Soldatova and Liakata 2007) and Argumentative Zoning-II scheme (AZ-II) (Teufel 1999). They concluded with a recommendation for merging the two schemes. Liakata (2010) further presented results of using the CoreSC annotation scheme for identifying negative and speculative statements in scientific papers.

2.2 Annotation of sentences with citations

Parallel to this line of investigation, there have been studies that have specifically examined sentences with citations in scientific articles. Two different streams of research can be identified in these studies, based on the approach involved: manual citation classification and automatic citation classification.

2.2.1 Manual citation classification

Baldi (1998) identified that ten different classification schemes were proposed between the years 1965 and 1980 and gently mocked this line of investigation as a cottage industry. These studies included those of Garfield (1965); Lipetz (1965); Hodges, (1972); Moravcsik and Murugesan (1975); Chubin and Moitra (1975); Spiegel-Rosing (1977); Lindsey and Lindsey, (1978); Oppenheim and Renn, (1978); Finney, (1979); and Frost (1979). The categories defined in these schemes characterized various reasons for citations and the methodology for classification involved manual examination of the passage with citations, also known as 'citation

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

context' using human judgment. Small (1982) called this line of investigation 'citation context analysis'. In recent times, the focus has shifted to directly learning from citing authors' motives for citations through surveys and interviews (Brooks 1985, 1986; Prabha 1986; Shadish *et al.* 1995; Case and Higgins 2000). White (2004) calls this stream of research 'citer motivations'.

2.2.2 Automatic citation classification

Studies have also focused on the automation of citation classification in scientific documents. These studies have either proposed a new or a modified scheme of citation classification in order to suit the automation process. Garzone (1997) developed an automated citation classifier with several components for carrying out citation classification. Nanba and Okumura (1999) proposed a simplified classification scheme of three categories based on the fifteen reasons identified by Garfield (1965) and developed a semi-automatic citation classification system for achieving multi-paper summarization.

Garzone and Mercer (2000) claim to have developed the first fully automatic citation classifier system. This extracted and processed only sentences with citations and used a pragmatic grammar for classification. Nanba, Kando, and Okumura (2000) used citation classification for achieving automatic generation of a review article. Pham and Hoffmann (2003) developed KAFTAN, the "Knowledge Acquisition Framework for TAsks in Natural language", which rapidly acquired cue phrases for classifying citations into four citation types. The authors compared the results of their classifier with the results of Nanba *et al.* (2000) and showed that their approach performed better.

Teufel, Siddharthan, and Tidhar (2006) presented an annotation scheme for citations involving twelve categories, based on the categories proposed by Spiegel-Rosing (1977), and used supervised learning experiments for automatically classifying citation sentences. Le, Ho, and Nakamori (2006) used graphical probabilistic models – Hidden Markov Models (HMMs) and Maximum Entropy Markov Models (MEMMs) – for extracting citation types from scientific articles. Radoulov (2008) carried out a study exploring automatic citation classification and redesigned Garzone's scheme of thirty-five categories from a perspective of usability and usefulness.

2.3 Motivation and rationale

Our work draws motivation from both types of studies discussed above. The objective of our study is to provide information services for the academic community using citation contexts, which requires a typology that describes scientific discourse along with different citation functions and a mechanism for identifying these functions. We explain in the following sections the rationale for defining a new set of context type definitions for sentences and also the reason for using sequential classification models for classification.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

2.3.1 Defining new set of context type definitions – the rationale

Though there are various annotation schemes readily available, it is not easy to use them as they are highly subjective. Baldi (1998) observed that citation classification typologies developed during 1960–1980 were designed in an ad hoc manner and virtually isolated from one another. White (2006) described the earlier classification schemes as idiosyncratic and difficult to code, resulting in difficulties in reusing them across disciplines. However, the importance of these classification systems cannot be ignored. Though critical, Baldi (1998) does not undermine the importance of these classification schemes and regards them as essential in understanding the contextual nature of citations. Schemes proposed by Teufel (1999) and Teufel *et al.* (2006) have achieved good inter-rater reliability measures indicating the usefulness of the schemes and several studies have extended these schemes successfully to suit different domains.

Guo *et al.* (2010) analyzed three annotation schemes and concluded that the selection of a scheme should be task oriented and the optimal scheme for use should depend on the level of detail required by the application at hand. The principal focus of our study is to develop a tool that identifies and presents citation contexts to users. Thus, the basic typology required for this task is a scientific discourse scheme that identifies contexts for sentences. Furthermore, the contexts for sentences with citations should represent finer details of citation functions. However none of the available annotation schemes provide such functionality. While some have specifically focused on defining different types of citation contexts, others that focus on scientific discourse do not consider finer details of citation functions. We therefore resorted to developing a new set of context type definitions for sentences, specifically defined for providing citation context based information services for the research community. The context type definitions are presented in Section 3.3. We also show in Section 3.6 how our definitions relate to the annotation schemes of Teufel (1999), which describe scientific discourse, and Teufel *et al.* (2006), which define various categories for citation functions.

2.3.2 Use of conditional random fields – the rationale

Most of the studies discussed above view the sentence classification task as a nonsequential classification problem, where a given sentence is considered independently from other sentences. However, sentences are rhetorically related to each other and the above studies capture these rhetorical relations mainly using cue phrases. It also needs to be noted that authors follow a sequential pattern in the writing process. Swales (1984) observed a sequential pattern in paragraphs of the introduction section in 50% of the articles, he surveyed. He noted that in these paragraphs, authors usually start with sentences that establish the research field and continue to report on the previous research. After this, sentences prepare the readers for the present research and continue on to introduce that research. Our earlier work identified a sequential pattern based on the generic rhetorical pattern in paragraphs of related work sections in the research articles (Angrosh, Cranefield, and Stanger 2010). Thus, a sequential pattern exists in research writing and the task

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

of sentence classification can be considered as a sequential classification task that can be addressed by using sequential classification models.

Several sequential classification models such as HMMs, MEMMs, and CRFs have been used for exploiting sequential patterns for improving classifier accuracy (Rabiner 1989; McCallum, Freitag, and Pereira 2000; Lafferty, McCallum, and Pereira 2001). Though HMMs are used for solving various sequential problems such as recognition of speech (Rabiner 1989), handwriting (Hu, Brown, and Turin 1996), gesture (Tanguay 1995; Elmezain *et al.* 2008), and part-of-speech tagging (Kupiec 1992), they are observed to suffer from two key problems (McCallum *et al.* 2000). First, HMMs do not provide for incorporating features, which allow for richer representation of observations, and second, the model follows a more traditional approach of employing a generative joint model in order to solve a conditional problem with given observations. McCallum *et al.* (2000) introduced MEMMs for solving these problems and showed that MEMMs achieved good performance compared to HMMs.

Though MEMMs performed better, Lafferty *et al.* (2001) identified 'label bias problems' in these models and introduced CRFs as a solution to this problem. CRFs have properties of both generative and classification models. As classification models, they are able to accommodate many statistically correlated features of the inputs and are trained discriminately, and as generative models they can trade off decisions at different sequence positions for achieving a globally optimally labeling (Sha and Pereira 2003). Lafferty *et al.* (2001) showed that CRFs performed better than HMMs and MEMMs on synthetic and natural language data.

Various studies have successfully used sequential classification models for sentence classification. Hachey and Grover (2005) tested various classifiers such as C4.5, Naive Bayes, Winnow, Support Vector Machines (SVMs), and Maximum Entropy Models (MEMs) for classifying sentences in the legal domain and concluded that the SVM classifier and the maximum entropy sequence tagger are suitable classifiers for the task. Hirohata *et al.* (2008) looked at the task of identifying sections in abstracts as a sequential classification task and used CRFs for achieving better accuracy compared to SVMs. Chung (2009) used CRFs and achieved good accuracy in classifying sentences in abstracts of randomized controlled trials. Kim *et al.* (2011) used CRFs for classifying sentences in medical abstracts and to develop a system for supporting evidence-based medicine.

In the light of these studies, we consider sentence classification as a sequential classification task, where the goal is to predict a label sequence $<Yi>$, given an input sequence $<Xi>$. This implies that the label for a given sentence may depend not only on the features of an individual sentence but also on the adjacent labels.

3 Context types for sentences in academic articles

The Oxford English Dictionary (Oxford University Press (OUP) 2010) defines the word 'context' as "the parts of something written or spoken that immediately precede and follow a word or passage and clarify its meaning". Thus, the context of a given word can be obtained by analyzing its surrounding words. However in the words of

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

the famous economist, Alfred Marshall "every word has many shades of meaning, and therefore needs to be interpreted by the context" (Marshall 2009: 43). Thus, in order to identify the context of a sentence, we may have to consider each word carefully, which could result in a complex process. However, a context type for a sentence can be defined by identifying key terms and phrases that contribute in defining its context. For example, a phrase such as "This paper shows" indicates that the author in the sentence is referring to the results of the paper discussed. Similarly a phrase such as "limitations of the study" could refer to research gaps. There could be exceptions, as in spite of the presence of these terms, the sentence may hold a different meaning. Therefore, we follow an approach of identifying key terms and phrases that contribute in defining the context of a sentence and also consider that the context type of a sentence is partly dependent on the context types of preceding and following sentences. The following sections explain the context types defined for sentences following this approach.

3.1 Methodology for identifying context types

In order to identify context types for sentences, we used the first 40 articles from a dataset of 70 articles taken from Lecture Notes in Computer Science (LNCS). As the key objective of our study is to develop a citation context based information system, we focused only on paragraphs that contained citations. Each of the sentences in these paragraphs was manually analyzed to identify the context associated with the sentence.

As a first step, we classified sentences into two categories: citation sentences and non-citation sentences. Citation sentences are those that have a reference to a published or unpublished source. Specifically, this is an expression in the sentence that points to an entry in the bibliographic references section of the article for the purpose of acknowledging the cited work. This expression can either be a numeric expression such as "[1]" or "[1, 2]" or author names used in the sentence for referring to the cited work. Non-citation sentences are those that do not have any such expressions. We focused on identifying contexts for sentences belonging to these two categories. The process of analyzing sentences resulted in developing a decision tree shown in Figure 1, which contains a set of questions for identifying the context associated with a sentence. The decision tree served as a tool for classifying sentences.

3.2 Categories of terms

While identifying the context type of a sentence, we also focused on identifying terms and phrases that contribute in defining the context of the sentence. The objective was to use these terms and phrases as features for our machine-learning experiments (see Section 5.3.3) Accordingly, we defined ten categories of terms that contribute to defining the context of sentences. Table 1 provides a description of each of these categories along with a listing of example terms.

The analysis of sentences following the above methodology resulted in identifying the following context types for citation and non-citation sentences.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

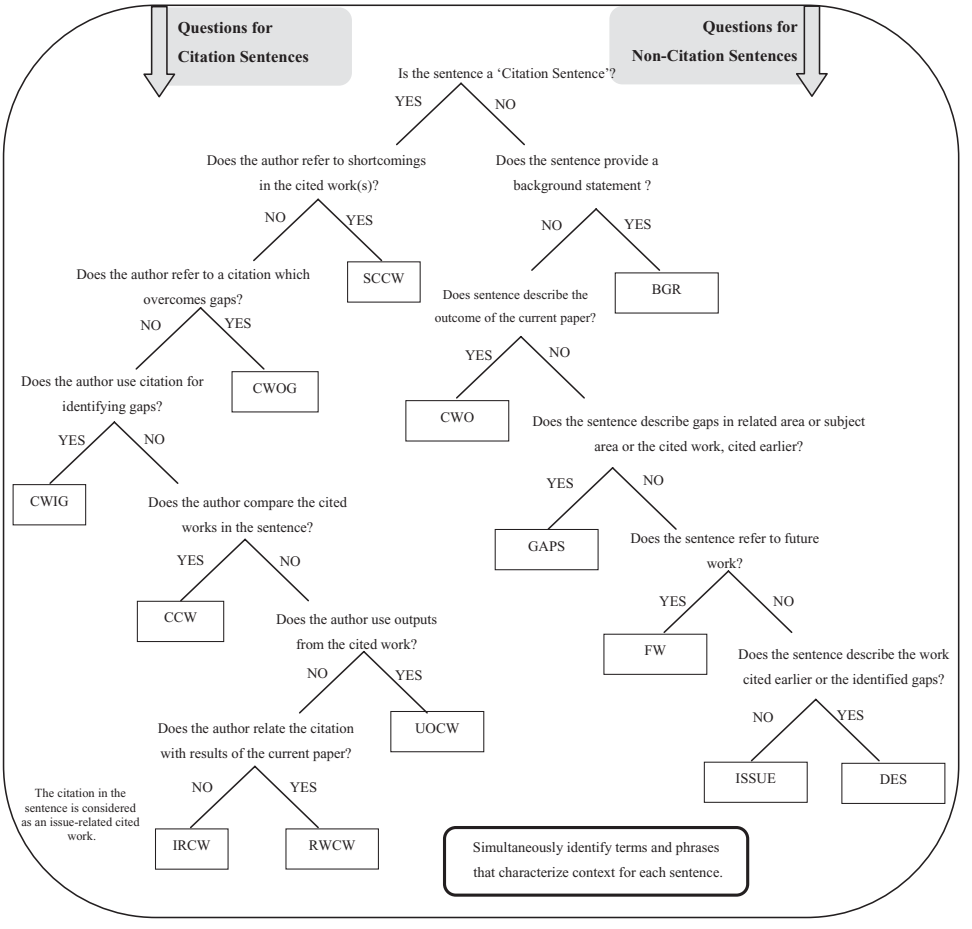


Fig. 1. Decision tree for identifying contexts of sentences.

3.3 Context types for citation sentences

The contexts associated with citation sentences are defined based on reasons for using the cited work in the current article. Accordingly, we distinguish between the contexts for citation sentences shown in Listing 1.

Listing 1. Context types for citation sentences

Context type definition	Description
Cited Work Identifies Gaps (CWIG)	The sentence uses cited work(s) for identifying gaps.
Cited Work Overcomes Gaps (CWOG)	The sentence discusses how the cited work(s) overcomes the identified gaps.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Uses Outputs from Cited Works (UOCW)	The sentence refers to using outputs of the cited work(s) in the work reported in the current article.
Results with Cited Work (RWCW)	The sentence relates the results of the article to the cited work(s).
Compare Cited Works (CCW)	The sentence compares different cited works.
Shortcomings in Cited Work (SCCW)	The sentence refers to shortcomings or limitations of the cited work(s).
Issue-Related Cited Work (IRCW)	The sentence cites other work(s) for other issues such as issues in research topics and the subject area discussed in the paper.

3.4 Context types for non-citation sentences

The contexts associated with non-citation sentences are defined based on the overall objective of the sentence. Listing 2 provides the definitions of various context types defined for non-citation sentences.

Listing 2. Context types for non-citation sentences

Context type definition	Description
Background (BGR)	The sentence provides a background statement.
Issues (ISSUE)	The sentence refers to issues identified or raised by the author.
Gaps (GAPS)	The sentence identifies gaps in general. These could be gaps in the research topics addressed or related to the current article, in works cited by the article, or in the current article itself.
Description (DES)	The sentence is a descriptive sentence, describing further a work cited earlier, methodology, gaps, issues, or background information.
Current Work Outcome (CWO)	The sentence refers to the results reported in the current article.
Future Work (FW)	The sentence refers to the future work proposed in the current article.

3.5 Framework for modelling the context types

In order to define relations between the above context types, we have proposed a framework shown in Figure 2. The framework considers that a paragraph with a citation is a set of reference areas, where a reference area is defined as the text

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 1. *Categories of terms and phrases in sentences*

	Category	Description	Example terms
1.	Comparing Terms (COM)	Terms or phrases that mention comparison studies.	Compared, evaluated
2.	Connecting Terms (CT)	Terms or phrases that indicate relations between sentences. These terms usually connect a sentence with its preceding sentence.	They, therefore, according to these, for this purpose, furthermore
3.	Extending Terms (EXT)	Terms or phrases that discuss extending the current work with the cited work.	Builds on previous work
4.	Future Work Term (FWT)	Terms or phrases that describe the future work of the paper.	Future work, we plan to extend, will be investigated
5.	Identifier Terms (IDT)	Terms or phrases that identify gaps or shortcomings in the related work or the cited work.	As shown, observations in, according to
6.	Issue Terms (ISSUE)	Terms or phrases that refer to issues raised or pointed by author.	Therefore, thus, hence, for this purpose
7.	Methodology Terms (MET)	Terms or phrases that describe the methodology adopted or followed in the paper.	We consider, we use, we assume
8.	Overcoming Gap Terms (OGT)	Terms or phrases that describe the characteristic of overcoming the identified gaps or shortcomings.	Enhanced, superior, promising, improved, better potential
9.	Result Terms (RES)	Terms or phrases that describe the results achieved either by the current paper or the cited paper.	We will show, we discover, we summarize
10.	Shortcoming Terms (SCT)	Terms or phrases that describe the shortcomings or gaps.	Nevertheless, performance suffers, perform poorly, are not studied

with a citation sentence and the surrounding sentences (Nanba and Okumura 1999). Several patterns can exist in reference areas that can involve one or more sentences defined in the framework.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

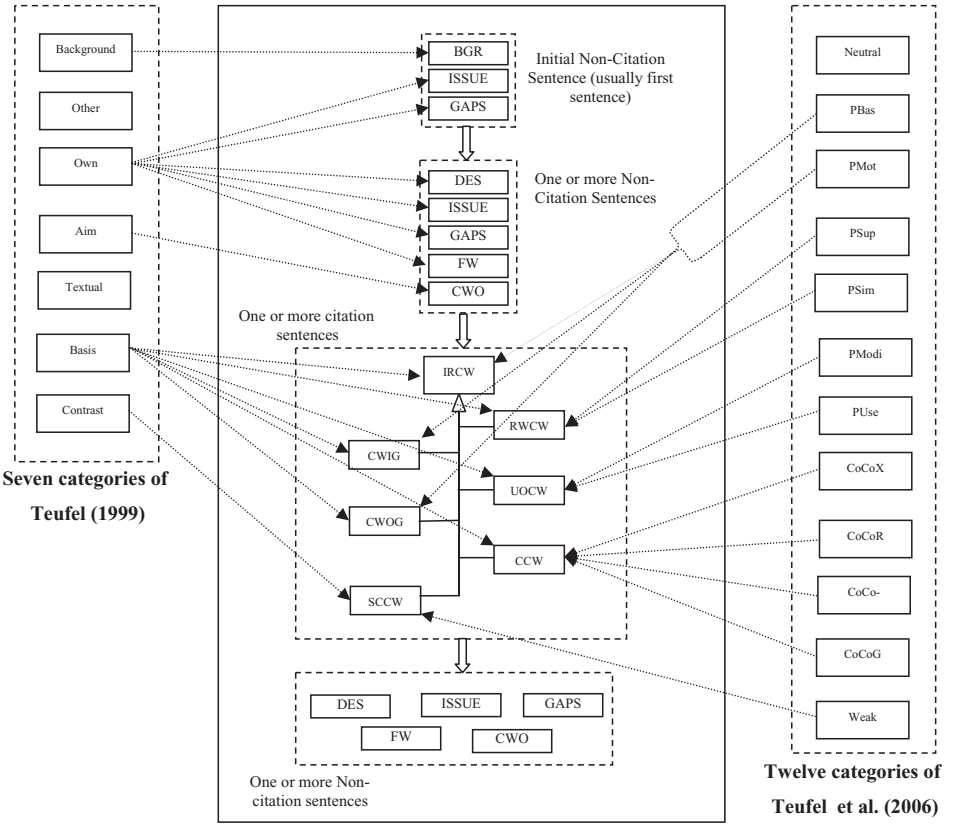


Fig. 2. Comparison of our annotation scheme with other schemes.

3.6 Comparison of our context types in relation with other schemes

We compare in this section the context types defined in our study with schemes proposed by Teufel and Moens (1999) and Teufel *et al.* (2006). The comparison is shown in Figure 2 and is described in the following sections.

3.6.1 Context types of sentences without citations

Some of the categories defined in our scheme can be directly compared to the categories proposed by Teufel and Moens (1999). For example, their class of Background is equivalent to the BGR class of our scheme. Similarly the categories Aim and Contrast, which refer to sentences describing research goals of the current article and weaknesses in cited works, are the same as CWO and SCCW in our scheme. However, their classes Own and Basis can be seen as the unions of classes of our scheme. The class Basis of Teufel and Moens (1999) consists of sentences that use cited work as a basis or starting point or gets support from them. While this is a broad definition, the classes of CCW, CWIG, CWOG, IRCW, RWCW, and

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

UOCW of our scheme are specific instances for using the cited work as a basis. Thus, a union of all these classes can be considered equivalent to the class Basis. Similarly, the class Own is equivalent to the union of classes of DES, ISSUE, FW, and GAPS, which describe aspects of the own work presented in the paper. We do not use the Textual and Other class, as we do not define anything similar to this class in our scheme. We further compare our classifier results with this scheme in Section 7.1.2.1.

3.6.2 Context types of sentences with citations

Teufel *et al.* (2006) have proposed a classification scheme for citation sentences. In this section, we compare our context type definitions for citation sentences with their scheme. The authors define the class Weak with sentences that refer to weaknesses of previous research. The class SCCW defined in our study is equivalent to the Weak class. The next four categories are used to describe comparisons or contrasts between the author's own and other work, which is equivalent to CCW in our scheme. The authors also define a set of categories for referring to statements that mention the active use of other works in the current article. Among them, the categories PBas and PMot identify the use of cited works as a basis or starting point or as a motivation factor in the current work, respectively. The categories IRCW, CWIG, and CWOG defined in our study represent these notions and thus, the union of these classes can be considered equivalent to the union of PBas and PMot. Similarly, the categories PSim and PSup are used to represent similarity and compatibility between the current work and the cited work. The category RWCW defined in our study signifies these notions and thus, a union of PSim and PSup is equivalent to RWCW. The categories PUse and PModi identify statements where the author uses and adapts or modifies tools, algorithms, or data from the cited work. The category UOCW signifies these characteristics. We do not identify any of our categories with Neutral defined in the other study. We further compare our classifier results with this scheme in Section 7.1.2.2.

4 Text preparation

The task of sentence classification involves various tasks for extracting structured information from unstructured text. These include (a) processing PDF documents to obtain usable content; (b) identifying section information for sentences; (c) sentence segmentation; (d) parsing sentences to identify terms and phrases; (e) generating features based on the presence of terms and phrases; (f) identifying keywords in sentences; and (g) parsing reference sections to obtain information about cited documents. In order to solve these tasks, we developed an information extraction and management system with various modules. The architecture of the developed system is shown in Figure 3. The key functions of the different modules are as follows.

The Download Manager uses XPath expressions and Java HTTP URLs for web scraping. The PDF Document Manager module is used for downloading PDF files, cropping PDF files to remove headers and footers, and converting PDF to

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

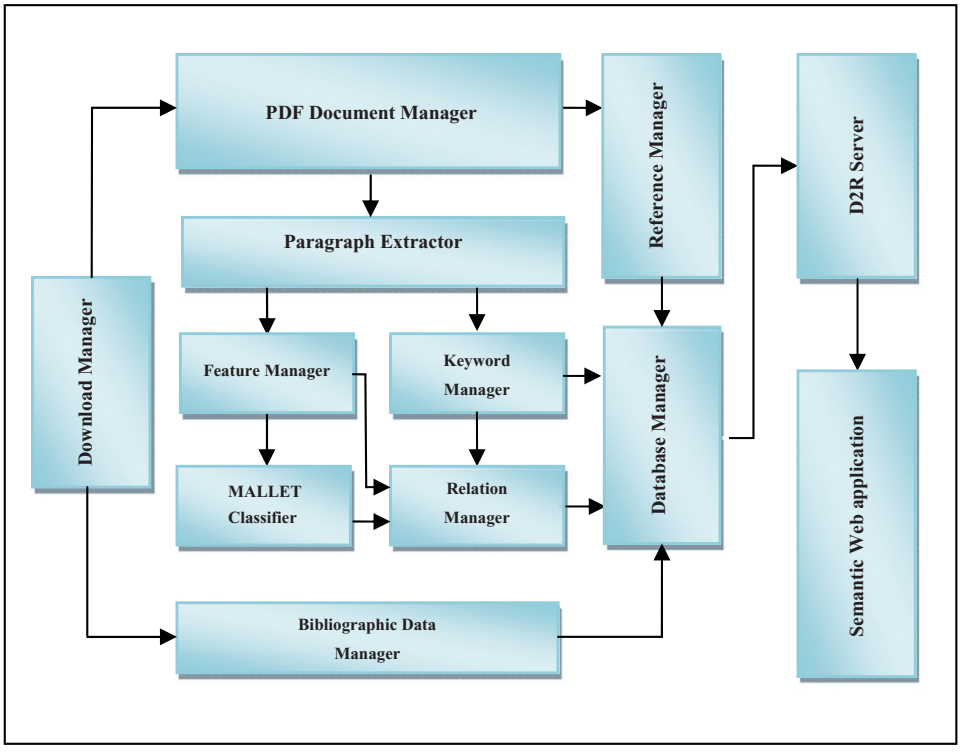


Fig. 3. (Colour online) System architecture of our information extraction system.

text and XML files. The Paragraph Extractor extracts paragraphs with citation sentences from text files along with section information and the Keyword Manager identifies keywords from citation sentences. The Feature Manager is responsible for parsing paragraphs and segmenting sentences. This module identifies the presence of key terms and phrases and generates features for each sentence. The MALLET Classifier labels paragraphs using the classifier model. The Relation Manager is used to relate results obtained from the Document Manager, the Feature Manager, and the Classifier.

The Reference Manager extracts references, segments references, identifies entities such as author names, the article title, and article source in each reference, and the Bibliographic Data Manager stores bibliographic metadata of research articles. The Database Manager stores data in a relational database management system and the D2R Server is used to generate RDF data from the relational database using a mapping file, configured according to the SENTCON ontology (Angrosh, Cranefield, and Stanger 2011). Finally, the resulting RDF data from the D2R server are used as input for the Semantic Web application.

5 Machine-learning experiments using sequential classifiers

In Section 2.3.2, we presented the rationale for using sequential classification models in our study. We report in this section the experiments carried out with CRFs

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

for achieving the task of automatic context identification. These experiments were performed for learning probabilities of sentence labels, given features that encode observed properties of sentences. This resulted in a trained CRF that we use for generating maximum likelihood sentence labels for new (unlabelled) paragraphs.

5.1 Conditional Random Fields

CRFs are undirected graphical models that define a single log-linear probability distribution over label sequences given an observation sequence (Lafferty *et al.* 2001). The structure of the graph in a CRF encodes independence relationships between labels and not the observations. This graphical structure facilitates a functional form of the distribution. This function combines several different terms known as *clique potentials* into a single product, in which each term forms a subset of the variables drawn from the full model.

The conditional probability of the labels gives the observations in a CRF with a linear chain structure (where the probability of each state y_i depends only on the probability of the state y_{i-1} and y_{i+1}, and the observed data sequence) factors according to the following equation (Vail 2008):

$$P(Y|X) = \frac{1}{Z_x} \prod_t \psi_t(y_{t-1}, y_t, X),$$

where X is the observation sequence, Y is the label sequence, and ψ_t is an arbitrary non-negative function. The normalization constant is computed by summing over all possible label sequences Y', which is tractable for linear chain structures using dynamic programming:

$$Z_x = \sum_{Y'} \prod_t \psi_t(y_{t-1}, y_t, X)$$

CRFs use a particular functional form for their 'clique functions' ψ_t:

$$\psi_t(y_{t-1}, y_t, X) = \exp(w^T f(y_{t-1}, y_t, X)),$$

where w is a real-valued weight vector and f is a vector of feature functions. The weights w are the model parameters that are estimated during the training phase.

5.2 Dataset and experimental setup, training, and evaluation

5.2.1 Dataset and experimental setup

We conducted experiments with a dataset of 1,000 paragraphs with citations extracted from seventy research articles chosen from LNCS. We followed the following steps in our experiments.

(1) Feature Definition – We manually examined sentences in paragraphs with citations in a training dataset of forty articles and manually labelled them according to the annotation scheme described in Section 3.3. The features defined for the sentences are described in Section 5.3.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

(2) Feature Selection – We performed experiments for feature selection and identified the optimal set of features. The feature selection process is described in Section 5.4.
(3) Training the Classifier – We trained the classifier model with the training dataset of forty articles using the feature set identified during feature selection. To this end, we conducted tenfold cross-validation experiments. Section 5.4 provides further details about training the classifier.
(4) Testing – Finally, we tested the developed classifier model on a test dataset of thirty articles. The results obtained with the test data are provided in Section 5.5.

For training the CRF model, we used MALLET (McCallum 2002), a Java-based package that provides an implementation of linear chain CRF algorithms for working with sequential data.

We computed precision, recall, and the *F*-score for measuring classification accuracy for each label. The *F*-score is computed as follows:

$$P = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|}; \quad R = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|}; \quad F = \frac{|2\mathrm{PR}|}{\mathrm{P} + \mathrm{R}},$$

where *P* represents precision, *R* represents recall, TP is the set of true positives, TN is the set of true negatives, and FP is the set of false negatives. A tenfold cross validation was performed for analyzing the performance of different feature sets on the training dataset. We describe in the following sections the results achieved in our experiments.

5.3 Feature definition

The following features were defined for sentences in our experiments.

5.3.1 Citation features

Citation features indicate whether a given sentence has a citation or not. The distinction is made based on the presence of a citation reference in the sentence. The application uses regular expressions for identifying the presence of citation reference expressions. In addition, references to citations can be made using the names of authors listed in the references section in the sentence without using an explicit reference. In such cases, the application looks for author names and terms such as '*et al.*', to decide about the status of the sentence.

5.3.2 Section features

Section features indicate the section of an article to which the sentence belongs. To define section features, we adopted a model where the content of an article is divided into three general blocks: introduction, body, and conclusion. Sections of an article with the titles 'Introduction', 'Related Work', 'Motivation', or 'Overview' are considered part of the Introduction block. Sections with the titles 'Conclusion', 'Conclusions and Future Work', or 'Future Work' are considered part of the

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 2. *Features defined for sentences*

Feature	Description
Citation features	
sentHasCitation	Sentence has citation.
prevSentHasCitation	Previous sentence has citation.
Block features	
sentSec=Intro	Sentence belongs to the Introduction block.
sentSec=Sub	Sentence belongs to the Body block.
sentSec=Conc	Sentence belongs to Conclusion block.
Term features	
sentHasTerm=CT	Sentence contains a connecting term or phrase.
sentHasTerm=SCT	Sentence contains a shortcoming term or phrase.
sentHasTerm=MET	Sentence contains a methodology term or phrase.
sentHasTerm=RES	Sentence contains a results term or phrase.
sentHasTerm=FWT	Sentence contains a future work term or phrase.
sentHasTerm=OGT	Sentence contains an overcoming gap term.
sentHasTerm=IDT	Sentence contains an identifier term.
sentHasTerm=EXT	Sentence contains an extending term.
sentHasTerm=COM	Sentence contains a comparing term.

conclusion block. Sections with other titles are considered under the Body block. It needs to be noted that the Related Work section in the article may appear anywhere in the article. Irrespective of its position, this section is considered part of the introduction block.

5.3.3 Term features

Term features for sentences are defined based on the presence of certain kinds of terms and phrases in the sentence that contribute to the context of the sentence. Table 1 lists the ten categories of terms and phrases that may be relevant to the context of the sentence. Table 1 also provides a description of each of these categories along with examples identified in each of them.

5.3.4 Normalization

Each sentence was normalized to a set of features based on the presence of the features defined above. For example, for a sentence that has a citation, a feature 'sentHasCitation' is created for the sentence. The features that are created for different cases are provided in Table 2. We developed Python scripts by using regular expressions for identifying and creating features for sentences.

5.4 Feature selection and training the classifier model

In order to identify the optimal set of features, we followed the backward selection method, where the performance of the classifier was tested on the test data by removing one feature from the dataset at a time. However, we did not evaluate

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 3. *Results of feature selection experiments*

Feature selected	CRF first-order			CRF first- and zero-order		
	CC	WC	A (%)	CC	WC	A (%)
Without ISSUE feature	186	73	71.81	191	68	73.74
Without SCT feature	160	99	61.77	193	66	74.51
Without RES feature	188	71	72.58	216	43	83.39
Without OGT feature	178	81	68.72	217	42	83.78
Without FWT feature	190	69	73.35	225	34	86.87
Without CT feature	194	65	74.90	226	33	87.25
Without MET feature	188	71	72.58	226	33	87.25
Without IDT feature	189	70	72.97	226	33	87.25
All features	194	65	74.90	230	29	88.80
Without prevSentHasCitation feature	190	69	73.35	231	28	89.18
Total number of sentences in test dataset: 259						

Note: CC – Correctly Classified; WC – Wrongly Classified; A – Accuracy.

the absence of the features – sentHasTerm=EXT and sentHasTerm=COM as these features occurred very few times in the training data and thus would not have impacted on the classifier model.

Thus, for this purpose, we divided the training dataset in the ratio of 9:1 and used the first part for training and the second for testing. We conducted experiments using two CRF structures – a first-order linear chain and a linear chain with additional zero-order features. The reason for using the second structure was due to the fact that our earlier experiments with related works sections show that a first-order linear chain performs poorly for classes that occur less often in the training data (Angrosh *et al.* 2010). A first-order CRF with additional zero-order features provides a 'back-off' prediction capability, thus increasing the accuracy for classes that appear less often. Table 3 provides details of the feature selection process.

As seen in Table 3, the removal of ISSUE features resulted in the maximum discrepancy by achieving only 73.74%. The use of all the features resulted in achieving a good accuracy of 88.80%. Interestingly, the removal of the feature prevSentHasCitation resulted in achieving the maximum accuracy and also a higher accuracy of 89.18% than the accuracy of 88.80% achieved using all the features. However, the difference between the two does not seem to be very significant. Thus, in order to further ascertain the role of the prevSentHasCitation feature, we conducted a tenfold cross validation on two training datasets: one that used all the features and the other without prevSentHasCitation feature.

In both the cases, we conducted experiments by using a CRF first-order linear chain and a CRF with additional zero-order features. The results of the classifiers are tabulated in Table 4 and the performance is graphically shown in Figure 4. In both cases, the first-order linear chain CRF with additional zero-order features performed better than the first-order linear chain, particularly for classes with a lower number of instances such as CCW and RWCW. Furthermore, the first-order linear chain CRF with zero order features scored almost the same for both training datasets with

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 4. *Results of the classifier for training dataset*

	All features						Without prevSentHasCitation feature					
	CRF first-order Accuracy: 89.97%			CRF first- and zero-order Accuracy: 91.70%			CRF first-order Accuracy: 90.01%			CRF first- and zero-order Accuracy: 91.58%		
Label	P	R	F	P	R	F	P	R	F	P	R	F
Non-citation sentences												
GAPS	0.94	0.96	0.95	0.95	0.95	0.95	0.91	0.96	0.93	0.95	0.95	0.95
DES	0.94	0.94	0.94	0.95	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.94
ISSUE	0.92	0.93	0.93	0.94	0.93	0.93	0.93	0.94	0.93	0.94	0.94	0.94
FW	0.96	0.63	0.76	0.97	0.86	0.91	0.95	0.60	0.74	0.91	0.86	0.89
BGR	0.87	0.87	0.87	0.86	0.87	0.87	0.88	0.87	0.87	0.86	0.86	0.86
CWO	0.85	0.85	0.85	0.84	0.89	0.86	0.84	0.84	0.84	0.86	0.89	0.88
Citation sentences												
IRCW	0.85	0.94	0.89	0.89	0.94	0.92	0.85	0.94	0.89	0.89	0.94	0.91
CWOG	0.90	0.72	0.80	0.88	0.88	0.88	0.87	0.69	0.77	0.88	0.88	0.88
UOCW	0.79	0.75	0.77	0.80	0.79	0.79	0.79	0.75	0.77	0.83	0.79	0.80
CWIG	0.83	0.59	0.69	0.79	0.73	0.76	0.90	0.64	0.75	0.74	0.69	0.71
SCCW	0.67	0.57	0.62	0.72	0.72	0.72	0.78	0.66	0.72	0.68	0.72	0.70
CCW	1.00	0.20	0.33	0.83	0.50	0.62	1.00	0.20	0.33	0.83	0.50	0.62
RWCW	0.40	0.35	0.37	0.76	0.51	0.61	0.73	0.35	0.47	0.80	0.54	0.65

Note: P – Precision; R – Recall; F – F-Score.

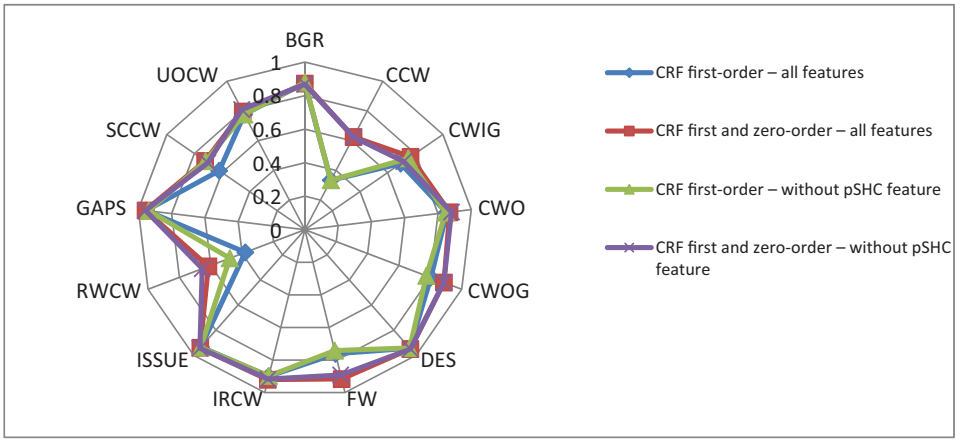


Fig. 4. (Colour online) Performance of classifiers for training dataset.

an accuracy of 91.70% and 91.58%, respectively. The difference between the two was very small. Therefore, in order to measure the statistical significance between the classifiers, we performed McNemar's test on the results.

The contingency tables and results of McNemar's test for both the datasets are shown in Figure 5. As seen, the *p*-value for McNemar's test for the dataset that had all the features is less than 0.05, indicating a significant difference between the

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Contingency table for training dataset with all features

Number of sentences misclassified by both	Number of sentences misclassified by 1
466	**255**
Number of sentences misclassified by 2	Number of sentences correctly classified by both
211	**4620**

McNemar's chi-squared = 3.9678, *df* = 1, *p*-value = 0.04638

Contingency table for training dataset without prevSentHasCitation feature

Number of sentences misclassified by both	Number of sentences misclassified by 1
468	**254**
Number of sentences misclassified by 2	Number of sentences correctly classified by both
214	**4618**

McNemar's chi-squared = 3.25, *df* = 1, *p*-value = 0.07142

Fig. 5. Contingency tables and results for McNemar's test.

two classifiers. The *p*-value for the other dataset was more than 0.07. Based on this statistical significance test, we decided to use the training dataset that included all the features for developing our classifier model and tested it on our test dataset.

5.5 Testing

After achieving the results described above, we trained a classifier model with the first-order linear chain with additional zero-order features, using the training dataset of forty articles, including all the features. We used this classifier model to test articles in the test dataset of thirty articles. Further, in order to examine the effect of newer terms on the classification accuracy, we divided the test dataset of thirty into three sets, with each set consisting of ten articles. We conducted the following two different sets of experiments on each of this dataset.

(1) Derive term features for each dataset using the terms obtained from the training dataset of forty articles.
(2) Derive term features for each dataset by appending terms from consecutive test datasets. The new terms obtained from test dataset 1 were added to the terms of the training dataset of forty articles for deriving term features for articles in test dataset 2. This step facilitated addition of newer terms to the term list for deriving term features.

The results of the classifier on the test dataset are shown in Table 5. As seen in the table, the classifier achieved an accuracy of 92.08%, 92.92%, and 90.01% with test datasets 1, 2, and 3, respectively, with inclusion of terms for deriving features. In comparison to this accuracy, the classifier achieved a lower accuracy of 89.03% and 89.15% for test datasets 2 and 3, respectively. The considerable difference is noted for the classes CWIG and GAPS (shaded rows in Table 5). The terms that are responsible for creating features for these classes belong to the 'Shortcoming Terms' category, which refers to gaps or shortcomings in the sentence. This indicates that it is necessary to develop a larger collection of terms belonging to this category for effectively creating this feature for presence of such terms. The average accuracy achieved for the full test dataset of thirty articles without addition of terms was

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 5. *Classifier results on test dataset*

	Test dataset 1 (ten articles)			Test dataset 2 (ten articles)						Test dataset 3 (ten articles)					
				Without addition of terms from test dataset 1			With addition of terms from test dataset 1			Without addition of terms from test datasets 1 and 2			With addition of terms from test datasets 1 and 2		
	A: 92.08%			A: 89.03%			A: 92.92%			A: 88.15%			A: 90.01%		
Label	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
BGR	1.00	0.8	0.88	0.81	0.81	0.81	0.83	0.81	0.82	0.81	0.77	0.79	0.84	0.80	0.82
CCW	0.66	1.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.50	0.66	1.00	0.50	0.66
CWIG	0.77	0.70	0.73	1.00	0.25	0.40	0.91	0.91	0.91	0.60	0.33	0.42	0.83	0.55	0.66
CWO	0.82	0.85	0.83	0.83	0.76	0.80	0.91	0.84	0.88	0.90	0.80	0.85	0.90	0.80	0.85
CWOG	0.75	1.00	0.85	0.84	0.64	0.73	0.93	0.82	0.87	0.78	0.57	0.66	0.64	0.57	0.61
DES	0.92	0.99	0.95	0.91	0.98	0.94	0.93	0.97	0.95	0.90	0.96	0.93	0.94	0.98	0.96
FW	1.00	0.66	0.80	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.7	0.82	1.00	1.00	1.00
IRCW	0.91	0.96	0.93	0.85	0.96	0.91	0.91	0.97	0.94	0.88	0.96	0.92	0.90	0.96	0.93
ISSUE	0.91	0.98	0.95	0.94	0.94	0.94	0.96	0.94	0.95	0.91	0.98	0.95	0.91	0.96	0.94
RWCW	1.00	0.37	0.54	1.00	0.33	0.50	1.00	0.33	0.50	0.57	0.8	0.66	0.57	0.80	0.66
GAPS	1.00	0.82	0.90	0.98	0.84	0.91	0.98	0.91	0.95	0.92	0.84	0.88	0.94	0.88	0.91
SCCW	0.83	0.55	0.66	1.00	0.73	0.84	0.88	0.78	0.83	0.60	0.65	0.62	0.61	0.69	0.65
UOCW	0.81	0.60	0.69	1.00	0.54	0.70	0.80	0.72	0.76	1.00	0.69	0.81	1.00	0.69	0.81

Note: P – Precision; R – Recall; F – F-Score.

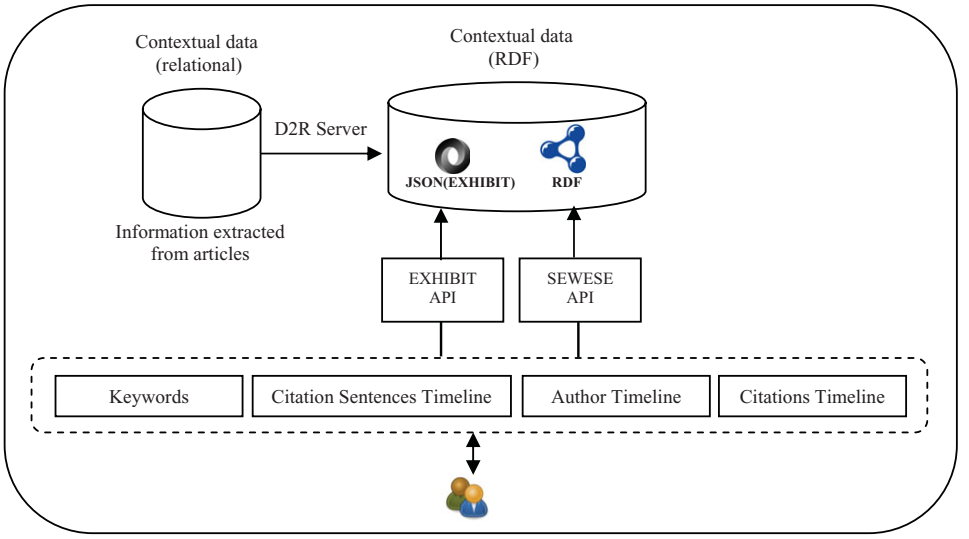


Fig. 6. (Colour online) Architecture of the Semantic Web application.

90.09% as against an average accuracy of 91.67% with inclusion of terms, resulting in a marginal difference of 1.58% between the two. Further, our analysis provided in Section 7 indicates that the number of additional terms added at each step for deriving term features became smaller indicating that the set of terms obtained from the dataset of seventy articles need not be continually extended to be applicable for newer data.

6 Citation context based information services

We describe in this section a Semantic Web application developed for providing citation context based information services, using the contextual data derived in the previous steps. The architecture of the application is shown in Figure 6. As seen in the figure, the D2R Server is used for converting the data in relational form into RDF. The RDF data is also converted into JSON (Exhibit) format using the Babel Service (Babel 2010) for using the data with the Exhibit API, a lightweight framework for publishing structured data on standard web servers (Huynh *et al.* 2007). The Timeline feature of the Exhibit API is used for browsing information with time as a reference point. The application also uses Semantic Web Server (SEWESE) tags and SPARQL queries for querying the RDF data. The application offers citation context based information services through various interfaces: the citation sentences timeline, author timeline, citations timeline, and keywords timeline, as explained in the following sections.

6.1 Browsing contexts of citation sentences in the article

The citation sentences timeline interface allows users to browse contexts of all citation sentences of an article. A screenshot of this interface is shown in Figure 7(a). As

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

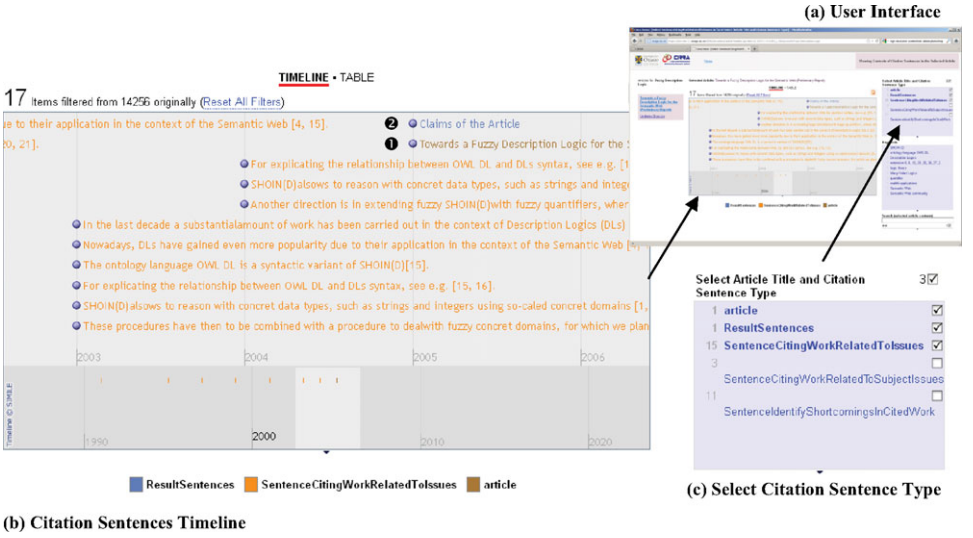


Fig. 7. (Colour online) Screenshot of citation sentences timeline.

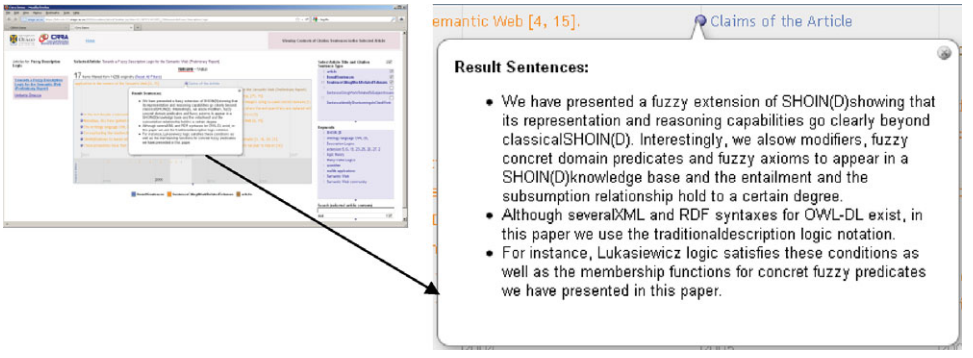


Fig. 8. (Colour online) Viewing result sentences of an article.

seen in the figure, the citation sentences of the article are displayed in the timeline. A zoomed-in image of the timeline alone, displaying citation sentences, is provided in Figure 7(b). The timeline feature provides an interactive interface, in which the user can move the timeline horizontally for viewing the displayed citation sentences on the timeline, placed according to the year of the cited work. As may be seen in Figure 7(b), the title of the selected article (indicated by the label ❶) and the 'Claims of the Article' (indicated by the label ❷) are displayed as links in the timeline. These links display the bibliographic details and the result sentences of the article, respectively. Figure 8 presents a screenshot where the user has selected to view the claims of the article and the result sentences identified by the classifier are displayed in the dialog box. There are also options for selecting specific types of citation sentences as shown in an enlarged version of the facet in Figure 7(c).

The application displays various related sentences for a citation sentence, following our Sentence Context Ontology (Angrosh *et al.* 2011). These include (1) Preceding

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

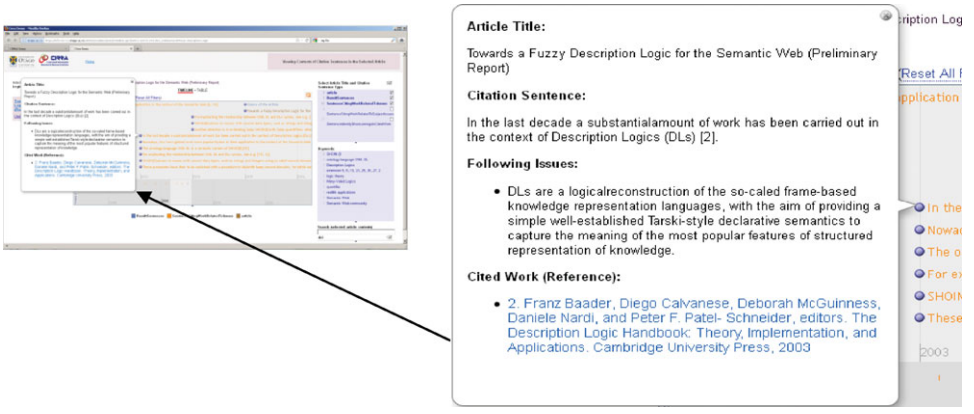


Fig. 9. (Colour online) Viewing associated sentences for the selected citation sentence.

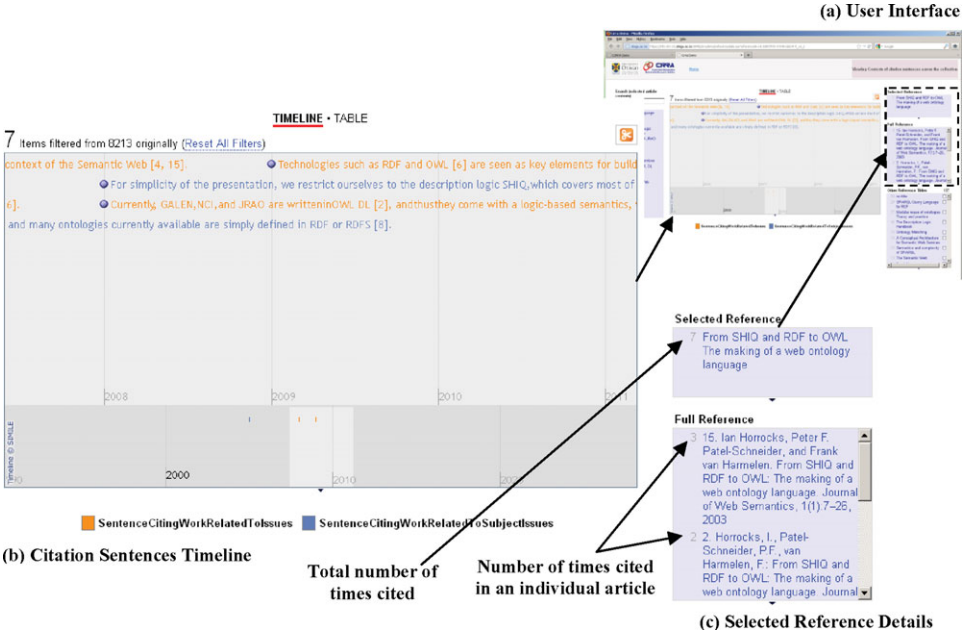


Fig. 10. (Colour online) Screenshot of the citations timeline.

Citation Sentence; (2) Preceding Issue Sentence; (3) Preceding Shortcoming Sentence; (4) Following Citation Sentence; (5) Following Description Sentence; (6) Following Shortcoming Sentence; and (7) Following Issue Sentence. In Figure 9, the application has identified the following issue sentences for the selected citation sentence. Further, the citation (reference) displayed in the popup dialog allows the user to navigate to the citations timeline for viewing the different contexts in which the citation is cited, as explained in the following section.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

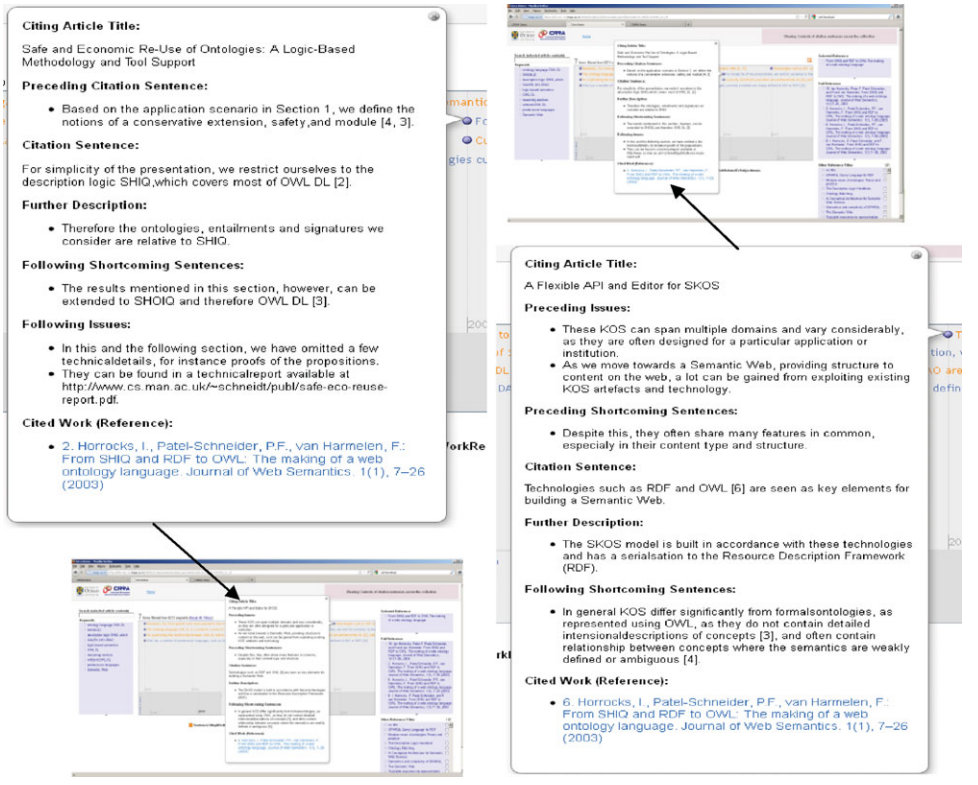


Fig. 11. (Colour online) Citations timeline – viewing details of different citation sentences.

6.2 *Browse contexts of citations*

The citations timeline interface allows users to view the contexts of citations across articles. A screenshot of this interface is shown in Figure 10(a). A zoomed-in image of the timeline alone, displaying citation sentences drawn from different articles, is provided in Figure 10(b). The facet on the right side (an enlarged version is provided in Figure 10(c)) provides information about the citation and the number of times it was cited in different articles.

Each of the sentences displayed on the citations timeline can be clicked to view its associated sentences. The popup dialog box displays the citing article title that has cited the citation along with the citation sentence and its associated sentences, as explained earlier in the case of the citation sentences timeline. The citations timeline helps users to view how different articles have cited a given work. An illustration of the usefulness of the citations timeline is provided in Figure 11. This figure provides a screenshot of viewing the contexts of the citations to an article from two citing articles. As can be seen, the dialog displays the title of the article citing the selected article along with the citation sentence and its associated sentences. This helps users in quickly viewing the contexts in which different articles have cited a given work. This feature gathers significance when a given work has several citations. It avoids the laborious process of examining the citation passages in different articles and provides an easier and efficient way of examining citation contexts.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

7 Discussion, analysis, and conclusion

We present in this section an analysis and comparison of our results with results of similar studies. With respect to supervised learning experiments, we discuss the use of different kinds of features in other studies and consider the usefulness of our approach. We also compare our classifier results with results of similar studies. With regard to the Semantic Web application, we discuss the citation-based information services provided by other tools and highlight the distinct features of our application.

7.1 Supervised learning experiments

7.1.1 Features

Previous studies have used various features for achieving the task of automatic citation classification. Prominent among them are cue phrases, which are linguistic indicators responsible for cohesion of text (Louwerse 2001 cited by Mancini 2005: 25) and features that identify the structure and location of a sentence. Teufel and Moens (1999), apart from using structural and location features, have used syntactic features that identify the tense, modality, and voice of the sentence, semantic features that capture the action, and content features that indicate the presence of keywords. The syntactic features were identified using part-of-speech (POS) information and semantic features were identified based on the presence of verbs using a manually developed lexicon. Heuristic evaluation was used to determine the tense and negation of a sentence.

Nanba *et al.* (2000) created rules using cue phrases and applied n-word gram analysis for achieving classification. Garzone and Mercer (2000) developed a pragmatic classifier that used cue phrases for building lexical matching and parsing rules. Pham and Hoffmann (2003) used human annotators for creating rules using cue phrases. Mizuta and Collier (2004b) identified multiple features through manual annotation for identifying zones in scientific articles in the biology domain. The authors observed the presence of similar features across zones and concluded that multiple features were essential in zone identification. Teufel *et al.* (2006) used both shallow and linguistic features, which included cue phrases, verb tense and voice, and sentence location. Hirohata (2008) used various words, word bigrams and a mixture of words and word bigrams as features for representing the content. They also used features for sentence location and to relate previous and following sentences to the current sentence.

The studies above indicate that cue phrases form an important feature in citation classification and are used for identifying rhetorical relations. Mercer and Marco (2003), after studying the frequency of cue phrases in citations, suggested that these phrases play an important role in citations. However, manual development of cue phrases and rules using them is a cumbersome and time-consuming task. Pham and Hoffman (2003) noted that certain cue phrases contradict each other, which results in a single sentence being matched by multiple cue phrases belonging to different classes, leading to conflicts. Identification of such conflicts is difficult and in order

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

to solve this problem, they automated the system to identify conflict sentences in the initial stage.

We followed an approach of identifying terms and phrases (explained in Table 1) that characterized the context of a sentence and used regular expressions for this task, which provided a concise and flexible means for matching terms and phrases in sentences. The features were derived based on the presence of terms and phrases. Generally, most of the categories have expressions that signify the context of the sentence, while the category of 'Connecting Terms' comprises pronouns that help in representing the relation between sentences. We also used features related to sentence location and citation status of the sentence. The citation features also contributed in describing relations between sentences.

In order to identify the increase in the number of terms in each category for newer paragraphs, we carried out an analysis of the terms. For this purpose, we divided the dataset of 1,000 paragraphs into four sets, each comprising 250 paragraphs and analyzed terms appearing in each of these sets. The results of our analysis are shown in Table 6.

As shown in Table 6 (column 1), the total number of terms identified in the first set of 250 paragraphs was 1,276. However, after removing duplicates there were 385 unique terms (column 2). The total number of terms identified in the second set, i.e., paragraphs 250 to 500 was 1,049 (column 3), of which there were 182 unique terms (column 4). Further, out of these 182 terms, 93 were already present in the first set (column 5). Thus, this set contributed ninety-two new terms (column 6). The average change in the number of terms from set 1 to set 2 was found to be 21% (column 8). The total number of unique terms in the set of the first 500 paragraphs was 477 (column 7). Similarly, the third set of 250 paragraphs contributed in adding twenty-five new terms (column 12) and the average change from the first 500 paragraphs to the third set was only 3% (column 14). The total number of unique terms in the first 750 paragraphs was 502 (column 13).

The fourth and the final set of 250 paragraphs contributed in adding only three new terms (column 18) to the total of 502, thus resulting in a total of 504 unique terms (column 19) for the dataset of 1,000 paragraphs. The average change from the first 750 paragraphs to the final set was 0.7%. This suggests that the increase in terms for new paragraphs is very low and the sets of terms gathered from these 1,000 paragraphs will not need to be continually extended to be applicable to new paragraphs.

7.1.2 Comparison of classifier results

We compare in this section the results achieved in our machine-learning experiments with similar studies.

7.1.2.1 Teufel and Moens (1999). We explained in Section 3.6.1 the comparison of our scheme with Teufel and Moens (1999). We present here a comparison of our results with their scheme. The authors present precision and recall measures for classification of sentences in seven categories. The comparison of our results with

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 6. *Details of number of terms identified in different categories*

	Paragraphs 1 to 250 (A)		Paragraphs 250 to 500 (B)						Paragraphs 500 to 750 (C)						Paragraphs 750 to 1,000 (D)					
	TT (1)	UT (2)	TT (3)	UT (4)	RT in Para 250 to 500 (5)	NT (6)	TUT in Para 1 to 500 (7)	PC (8)	TT (9)	UT (10)	RT in Para 500 to 750 (11)	NT (12)	TUT in Para 1 to 750 (13)	PC (14)	TT (15)	UT (16)	RT in Para 500 to 750 (17)	NT (18)	TUT in Para 1 to 1,000 (19)	PC (20)
COM	13	9	3	2	1	1	10	12	6	2	2	0	10	0	6	4	4	0	10	0
CT terms	488	49	418	30	30	1	50	2	315	28	28	0	50	0	339	31	31	0	50	0
EXT	04	04	0	0	0	0	4	0	0	0	0	0	4	0	0	0	0	0	4	0
FWT terms	36	26	16	8	3	5	31	20	4	2	1	1	32	4	11	6	5	1	32	4
IDT terms	123	29	90	22	7	15	44	52	80	19	16	3	47	7	71	15	15	0	47	0
MET terms	188	75	77	15	17	0	75	0	61	14	14	0	75	0	67	15	15	0	75	0
OGT terms	75	30	84	15	1	14	44	46	52	15	13	2	46	5	71	15	15	0	46	0
RES terms	134	67	115	34	18	16	83	23	57	25	20	5	88	7	77	23	22	1	89	2
SCT terms	215	96	246	56	16	40	136	42	136	39	25	14	150	11	159	26	25	1	151	1
	1,276	**385**	**1,049**	**182**	**93**	**92**	**477**		**711**	**144**	**119**	**25**	**502**		**801**	**135**	**132**	**3**	**504**	
					Average change (%)			**21**			Average change (%)			**3**			Average change (%)			**0.7**

Note: TT – Total Terms; UT – Unique Terms; RT – Repeated Terms; NT – New Terms; TUT – Total Unique Terms; PC – Percentage Change.

that of Teufel and Moens (1999) and the confusion matrix of the combined test data for different classes is provided in Table 7. As seen in Table 7, the results achieved in our study score better for all classes compared to their results.

7.1.2.2 Teufel et al. (2006). In Section 3.6.2, we presented a comparison of the context types defined in our study for citation sentences with the scheme of Teufel *et al.* (2006). The comparison of our results against the results achieved by them is shown in Table 8. As may be seen, our results score better than the results obtained by them.

7.2 Citation context based information system

Some of the popular search engines used by the academic community for searching scholarly literature include Google Scholar, Scopus, ScienceDirect, PubMed, and CiteSeerX. While Google Scholar provides search across different disciplines, Scopus provides search services in sciences and social sciences. ScienceDirect is used to search articles in sciences, and PubMed, and CiteSeerX cater to specific disciplines such as medicine and computer and information science. The prominent citation-related services provided by these search engines include citation statistics that provide information about the citing articles and citation links for accessing the citing articles. They further facilitate identifying related documents. While this information helps in identifying the number of cited works for a given article, they do not provide contexts in which the cited articles have used the article. The only way to learn these contexts is to read all cited documents, which is a labourious process.

The only search engine that provides citation context services is CiteSeer, which is a popular scientific literature digital library and search engine that automatically crawls and indexes scientific documents in the field of computer and information science (Li *et al.* 2006). Recently, CiteSeer has been upgraded with a new architecture and data model called CiteSeerX in order to overcome scaling problems in handling more documents, more features, and more users. CiteSeerX uses Autonomous Citation Indexing (ACI) (Lawrence, Giles, and Bollacker 1999) that automatically creates a citation index from literature in electronic format and provides additional citation-related services such as current awareness and community features that complement commercial citation indices such as SCI. ACI systems locate articles, extract citations, and identify citations to the same article that occur in different formats and also identify the context of citations in the body of article. However, none of the systems described above provide information about the context in which a citation is used by the citing articles. In comparison to the above systems, the following are the distinct features of our application.

7.2.1 Distinct features of our application

7.2.1.1 Viewing context types of citations in an article. Currently, there are no tools that enable users to view contexts of citations in an article. Our application enables users to view the use of citations in an article by presenting their context types (as

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 7. *Comparison of our results with Teufel and Moens (1999)*

Teufel and Moens (1999)		Our study					Confusion matrix for tests datasets 1–3												
Class	F	F	Class	P	R	F		BGR	CCW	CWIG	CWO	CWOG	DES	FW	IRCW	ISSUE	RWCW	GAPS	SCCW	UOCW	TOTAL
Background	0.38	0.89	BGR	0.89	0.80	0.84	BGR	94					20		2			1			117
Other	0.44	...	NOT USED																		
Own	0.86	0.94	DES U	0.93	0.98	0.95	DES	3			1		575		3	1		1			584
			ISSUE U	0.93	0.96	0.95	ISSUE	1					4		1	190					196
			GAPS U	0.97	0.88	0.92	GAPS	6		2	1	1	8			9		213	2		242
			FW	1.00	0.97	0.96	FW							12		1					13
Aim	0.52	0.82	CWO	0.87	0.82	0.85	CWO				63	4	6				3				76
Textual	0.61	...	NOT USED																		
Contrast	0.60	0.75	SCCW	0.73	0.70	0.72	SCCW					3			8			4	36		51
Basis	0.38	0.74	IRCW U	0.91	0.97	0.94	IRCW	1			2		1		365	3			1	3	376
			CWIG U	0.85	0.74	0.79	CWIG			23					3				5		31
			CWOG U	0.77	0.73	0.75	CWOG			2		31			3				5	1	42
			RWCWU	0.72	0.50	0.59	RWCW		1		5				2		8				16
			UOCW	0.86	0.66	0.75	UOCW					1	1		11					26	39
			CCW	0.75	0.60	0.66	CCW		3						2						5

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Table 8. *Comparison of our results with Teufel and Moens (2006)*

Teufel *et al.* (2006)		Our study										
				Confusion matrix of combined test datasets 1–3								
Class	F	F	Class		CCW	IRCW	CWIG	CWOG	RWCW	UOCW	SCCW	TCS	SC-NCS
Weak	0.60	0.72	SCCW	SCCW		8		3			36	51	4
CoCoGM CoCoR0 CoCo- CoCoXY	0.60	0.66	CCW	CCW	3	2						5	...
PBas PMot	0.58 0.69	0.82	IRCW U CWIG U CWOG	IRCW		365				3	1	376	7
				CWIG		3	23				5	31	...
				CWOG		3	2	31		1	5	42	...
PUse						11		1		26		39	1
PModi	0.50	0.75	UOCW	UOCW									
PSim	0.48												
PSup	0.47	0.59	RWCW	RWCW	1	2			8			16	5
Neutral	0.86	not used											

Note: TCS – Total Citation Sentences; SC-NCS – Sentences classified as Non-Citation Sentences.

discussed in Section 6.1). This helps in distinguishing between different citations in the article. For example, if the user is interested in seeing works cited by the article, for which the author has identified gaps, or has used them as a basis or a starting point, the application provides a classification of citation sentences according to different context types. This can be used to browse these specific citations. In the absence of such a feature, the user would be required to read the full article in order to identify the contexts of different citations. Further, it would also be difficult to keep track of the use of citations as the user progresses to read the article. Our application provides a useful feature in this direction.

7.2.1.2 Viewing contexts of citing articles. The application provides a useful feature in identifying the contexts of citing articles given a target article (as discussed in Section 6.2). As discussed earlier, the present available tools provide citation link services and do not help in identifying the context of a citation. Our application differs from these tools by providing the contextual information about the citation. In the absence of such a feature, the user has to go through a laborious process of identifying the context in various citing articles. Our application provides a valuable feature in this direction by extracting only the citation paragraphs from the citing articles and presenting them to the user along with their context, which helps in easier evaluation of the target article. An example use of this facility would be a researcher searching for citations to one of his articles where the cited author has used the article for identifying or overcoming gaps in the related research area or identified gaps in the article itself. The feature provided by our application helps to learn these contexts and understand the contexts in which an individual's article is cited.

7.2.1.3 Sketching the intellectual lineage for topics. The interface provides a keywords timeline (not discussed in Section 6) that allows users to browse citation sentences for a specific topic drawn from different articles, published over a period of time. This allows users to see how different authors have addressed a specific topic over a period of time and aids in sketching the intellectual lineage for ideas.

7.2.1.4 Viewing contexts of author works. The application also provides an author timeline (not discussed earlier) that displays citation sentences drawn from different articles belonging to the specified author. Each of these citation sentences can display its associated sentences and the source article. This interface helps users to browse an individual author's work cited in various contexts in different articles. This timeline also provides a ready reference to the works authored by the selected author, by displaying their titles and links for accessing the respective article.

7.3 Conclusion

We presented in this paper our research on identifying contexts associated with sentences in research articles and using this information for providing intelligent information services for the research community. To this end, we defined various

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

context types for sentences and compared our contextual definitions with other studies. We described machine-learning experiments on automatic context identification for sentences and provided an overview of an application developed for providing citation context based information services. We have also compared our classifier results with those from prior work and have shown that our results score well in comparison. The focus of this research has been not to replace the current existing information search systems but to provide additional services that complement the existing services. The citation context based information services discussed in this paper can be integrated with current information providing systems for designing useful information services for researchers. The different additional useful information services that can be provided were discussed in this paper. Our future work focuses on conducting an inter-rater reliability study for validating the objectivity of our chosen labels for sentences and also an evaluation study for assessing the usefulness of the web application.

References

Angrosh, M. A., Cranefield, S., and Stanger, N. 2010. Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries. In J. Hunter et al. (eds.), *Proceedings of the 2010 Joint Conference on Digital Libraries*, pp. 293–302. New York: ACM.

Angrosh, M. A., Cranefield, S., and Stanger, N. 2011. Contextual information retrieval in research articles: semantic publishing tools for the research community. *The Information Science Discussion Paper Series Number 2011/06*, Department of Information Science, Dunedin, University of Otago.

Baldi, S. 1998. Normative versus social constructivist processes in the allocation of citations: a network-analytic model. *American Sociological Review* **63**(6): 829–46.

Brooks, T. A. 1985. Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science* **36**(4): 223–9.

Brooks, T. A. 1986. Evidence of complex citer motivations. *Journal of the American Society for Information Science* **37**(1): 34–6.

Buckingham Shum, S. J., Uren, V., Li, G., Sereno, B., and Mancini, C. 2007. Modelling naturalistic argumentation in research literatures: representation and interaction design issues. *International Journal of Intelligent Systems* **22**(1): 17–47.

Case, D. O., and Higgins, G. M. 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science* **51**(7): 635–45.

Chubin, D. E., and Moitra, S. D. 1975. Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science* **5**(4): 423–41.

Chung, G. Y. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making* **9**(10): 1–13.

Elmezain, M., Al-Hamadi, A., Appenrodt, J., and Michaelis, B. 2008. A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory. In *Proceedings of the 19th International Conference on Pattern Recognition*, pp. 1–4. Tampa Florida: IEEE.

Finney, B. 1979. *The Reference Characteristics of Scientific Texts*. Master's thesis. London: The City University of London.

Frost, C. O. 1979. The use of citations in literary research: a preliminary classification of citation functions. *The Library Quarterly* **49**(4): 399–414.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Gaillard, J. 2008. The characteristics of R and D in developing countries: measuring R & D in developing countries, the UNESCO Institute of Statistics (UIS), March 2008. http://www.uis.unesco.org/template/pdf/S&T/Gaillard_final_report.pdf

Garfield, E. 1965. Can citation indexing be automated? In M. E. Stevens et al. (eds.), *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings*, vol. 1, pp. 189–92. Washington: National Bureau of Standards Miscellaneous Publication.

Garzone, M. A. 1997. *Automated Classification of Citations Using Linguistic Semantic Grammars*. MSc thesis. London: University of Western Ontario.

Garzone, M., and Mercer, R. E. 2000. Towards an automated citation classifier. In H. Hamilton, and Q. Yang (eds.), *Canadian AI 2000*, pp. 337–46. Lecture Notes in Artificial Intelligence, vol. 1822. Berlin: Springer-Verlag.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., Sun, L., and Stenius, U. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp. 99–107. Uppsala, Sweden: Association of Computational Linguistics.

Hachey, B., and Grover, C. 2005. Sequence modelling for sentence classification in a legal summarisation system. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, pp. 292–6. New York: ACM.

Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, Hyderabad, India, pp. 381–8.

Hodges, T. L. 1972. *Citation Indexing: Its Potential for Bibliographic Control*. PhD thesis. Berkeley: University of California.

Hu, J., Brown, M. K., and Turin, W. 1996. HMM based on-line handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(10): 1039–45.

Kim, S. N., Martinez, D., Cavedon, L., and Yencken, L. 2011. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics* **12**(Suppl 2):S5: 1–10.

Kupiec, J. 1992. Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language* **6**: 225–42.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–9. Williamstown, MA, USA: Morgan Kaufmann.

Langer, H., Lüngen, H., and Bayerl, P. S. 2004. Text type structure and logical document structure. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation - DiscAnnotation '04*, pp. 49–56, Barcelona, Spain: ACL.

Lawrence, S., Giles, C. L., and Bollacker, K. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer* **32**(6): 67–71.

Le, M. H., Ho, T. B., and Nakamori, Y. 2006. Detecting citation types using finite-state machines. In *PAKDD 2006*, LNCS 3918, pp. 265–74, Berlin, Heidelberg: Springer-Verlag.

Li, H., Councill, I., Lee, W. C., and Giles, C. L. 2006. CiteSeerX: an architecture and web service design for an academic document search engine. In *WWW 2006*, pp. 883–4. New York: ACM.

Liakata, M. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 1–4, Uppsala, Sweden.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association.

Lindsey, D., and Lindsey, T. 1978. The outlook of journal editors and referees on the normative criteria of scientific craftsmanship: viewpoints from psychology, social work, and sociology. *Quality and Quantity* **12**: 45–62.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Lipetz, B. 1965. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation* **16**(2): 81–90.

Marshall, A. 2009. *Principles of Economics*, 8th ed. New York: Cosimo.

McCallum, A. K. 2002. MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu

McCallum, A., Freitag, D., and Pereira, F. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 591–8. Stanford, CA: Morgan Kaufmann.

Mizuta, Y., and Collier, N. 2004a. An annotation scheme for a rhetorical analysis of biology articles. In *Proceedings of the Fourth International Conference on Language Resource and Evaluation, (LREC 2004)*, Lisbon, Portugal.

Mizuta, Y., and Collier, N. 2004b. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, (JNLPBA' 04)*, pp. 29–35. Geneva, Switzerland: ACL.

Moravcsik, M. J., and Murugesan, P. 1975. Some results on the function and quality of citations. *Social Studies of Science* **5**(1): 86–92.

Nanba, H., Kando, N., and Okumura, M. 2000. Classification of research papers using citation links and citation types: towards automatic review article generation. In *Proceedings of the 11th SIG Classification Research Workshop, Classification for User Support and Learning,* pp. 117–34, Chicago, USA.

Nanba, H., and Okumura, M. 1999. Towards multi-paper summarization retrieval of papers using reference information. In T. Dean (ed.), *IJCAI*, pp. 926–31. Stockholm, Sweden: Morgan Kaufmann.

Oppenheim, C., and Renn, S. P. 1978. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science* **29**(5): 227–31.

Oxford University Press (OUP). 2010. *Oxford Dictionary of English*. A. Stevenson (ed.). Oxford, UK: Oxford University Press.

Pham, S. B., and Hoffmann, A. 2003. A new approach for scientific citation using cue phrases. In T. D. Gedeon, and L. C. C. Fung (eds.), *Australian Joint Conference in Artificial* Intelligence, pp. 759–71. Berlin: Springer-Verlag.

Prabha, C. G. 1986. Some aspects of citation behavior: a pilot study in business administration. *Journal of the American Society for Information Science* **34**(3): 202–6.

Rabiner, L. R. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2): 257–86.

Radoulov, R. 2008. *Exploring Automatic Citation Classification.* MSc thesis. Ontario: University of Waterloo.

Research4Life. (2009). *Research output in developing countries reveals 194% increase in five years.* Program Manager. http://www.research4life.org/Documents/Increase_in_developing_country_research_output.pdf

Sha, F., and Pereira, F. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – (NAACL'03)*, pp. 134–41. Edmonton, Canada: Association for Computational Linguistics.

Shadish, W. R., Tolliver, D., Gray, M., and Sen Gupta, S. K. 1995. Author judgements about works they cite: three studies from psychology journals. *Social Studies of Science* **25**(3): 477–98.

Shatkay, H., Pan, F., Rzhetsky, A., and Wilbur, W. J. 2008. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics* **24**(18): 2086–93.

Small, H. 1982. Citation context analysis. In P. Dervin, and M. J. Voigt (eds.), *Progress in Communication Sciences 3*, pp. 287–310, Norwood, NJ: Ablex.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press

Soldatova, L., and Liakata, M. 2007. An ontology methodology and CISP - the proposed core information about scientific papers. JISC Technology and Standards Watch. Aberystwyth: The University of Wales. http://ie-repository.jisc.ac.uk/137/1/ReportCISP.pdf

Spiegel-Rosing, I. 1977. Science studies: bibliometric and content analysis. *Social Studies of Science* **7**(1): 97–113.

Swales, J. 1984. Citation analysis and discourse analysis. *Applied Linguistics* **7**(1): 39–56.

Tanguay, D. O. 1995. *Hidden Markov Models for Gesture Recognition*. MSc thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Boston.

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

Teufel, S., and Moens, M. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. In *ACL Workshop - Towards Standards and Tools for Discourse Tagging*. Maryland, USA: ACL.

Teufel, S., Siddharthan, A., and Tidhar, D. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp. 103–10. Sydney, Australia: ACL.

White, H. D. 2004. Citation analysis and discourse analysis revisited. *Applied Linguistics* **25**(1): 89–116.

Wilbur, W. J., Rzhetsky, A., and Shatkay, H. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* **7**: 356.

https://doi.org/10.1017/S1351324912000277 Published online by Cambridge University Press