Conditional Random Field Based Sentence Context Identification: Enhancing Citation Services for the Research Community

Angrosh M.A.  Stephen Cranefield  Nigel Stanger
Department of Information Science, University of Otago, Dunedin, New Zealand
(angrosh, scranefield, nstanger) @ infoscience.otago.ac.nz

Abstract

Academic publishers' full text databases are an important part of the deep Web for researchers and a potentially valuable resource for automated extraction of scientific knowledge. Recently, some major publishers have provided Web APIs for accessing their article databases, thus allowing the development of Web applications to mine these resources. However, the task of knowledge discovery from academic articles, particularly with citations, remains a challenge. We present in this paper our research work taken up for identifying contexts associated with sentences in academic articles and use of this information to provide information services for the research community. To this end, we propose an annotation scheme for sentences in academic articles. We also describe our experiments with conditional random fields for sentence classification. Finally, we present CitContExt—a citation context extraction application developed based on the techniques discussed above.

Keywords: Sentence Context Identification, Conditional Random Fields, Citation Context

1 Introduction

Academic publishers' full text databases are an important part of the deep Web for researchers and a potentially valuable resource for automated extraction of scientific knowledge. In particular, there are a number of prominent online academic databases that provide Web-based access to metadata about the full text of academic publications. One valuable service provided by some of these databases is the ability to search and browse the citation links between papers. Citations play an important role in research writing, as they serve to acknowledge the intellectual lineage of ideas and allow cited claims to be verified.

However, the available citation-related query services are limited to following citation links to locate cited articles, and to provide information about the number of articles citing an article.

Within articles, citations are presented in sentences, and these sentences along with the adjacent sentences provide a context that implicitly defines the purpose of the citations, e.g. to identify prior work that the citing article's author believes has gaps. It would be of great benefit to researchers and librarians if the services provided by online academic databases could be extended to extract this implicit information and support its use for querying, browsing and the display of statistics. However, there are various challenges that must be addressed to provide such contextual information services.

A major challenge in providing such knowledge-based services is the difficulty in handling the distinct characteristics of language and thought implicit in the research content. Another significant challenge is the form in which the research content is available in the online environment. Research articles are usually published as PDF documents on the Web, which makes it difficult to obtain reusable content from these documents, particularly at runtime. The formatting and style of research articles forms another barrier in providing contextual services. It is often hard to handle different formatting styles of research articles.

Interestingly, in recent times, some of the major publishers have provided Web APIs for accessing their article databases, thus allowing the development of Web applications to mine these resources (Sciverse 2011; Springer 2011). These Web APIs can be used to access reusable research content that have uniform style and formatting solving the issues related to content discussed above. Thus, the missing component is a knowledge extraction tool that identifies contexts associated at sentence level in research articles. Identifying this need, this study is taken up for developing a citation context extracting tool that can be integrated with Web APIs to provide knowledge-based services for the research community. Our initial work focussed on defining and modelling contexts of sentences in related work sections of scientific articles (Angrosh et al. 2010a; Angrosh et al. 2010b). These experiments were further extended to cover all citations in the full text and this paper describes these results.

One of the requisites to provide such services is a set of context type definitions that signify different contexts associated with sentences. Though there are a few annotation schemes, which we discuss in Section 2, we found it difficult to reuse these schemes for our application and resorted to define a new set of context type definitions for sentences. Further explanation of the rationale for this is provided in Section 2.1 and our set of context type definitions are provided in Section 3. Section 3 also discusses the methodology followed for deriving these definitions. The process of context definition also resulted in developing a framework for modelling these contexts. This is presented in Section 3.3.

At the core of a citation context based application lies text mining or data mining techniques that are used for
automatically identifying the contexts associated with sentences. There have been studies that have explored the possibility of automatically identifying contexts associated with sentences in research articles. Some of these studies are identified in Section 2. We view sentence classification as a sequential problem and employ supervised learning methods using sequential classification models such as Conditional Random Fields (CRFs) for sentence classification. Further rationale for using sequential classification models is provided in Section 2.2 and the description of the experiments carried out with CRFs is provided in Section 4.

Another integral part of a citation related application is the user interface that can be effectively used for consuming the provided information services. To provide such an interface, we developed CitContExt – a citation context extracting tool that can be hosted on the Elsevier Sciverse Applications platform. CitContExt demonstrates the possibility of providing citation context based information services at runtime by using Web APIs. It uses the Sciverse Content and Scopus APIs to obtain the full text of articles and metadata about the articles they cite, performs feature extraction on sentences, and applies a pre-trained sentence context classifier that was generated from an online document collection using the supervised machine learning techniques mentioned above. Section 5 describes the architecture and the information services provided by CitContExt. In Section 6, we discuss use cases of our application and conclude this paper in Section 7 with pointers to our future work.

2 Related Work

Over the years, the task of identifying contexts associated with sentences in research articles has received significant attention. Much of the focus has been on creating text summarization. Teufel and Moens identified the context of sentences based on the notion of argumentation (Teufel & Moens 1998). Hara and Matsumoto applied sentence classification techniques to the clinical trial abstracts for achieving summarization (Hara & Matsumoto 2005). Zaanen et al. used machine learning techniques and regular expressions for classifying sentences (Zaanen et al. 2005).

There has also been specific focus on identifying contexts associated with sentences that have citations. These characterize the reasons for using cited works in the article. As early as 1965, Eugene Garfield noted fifteen different reasons for authors to cite other people’s work (Garfield 1965). Moravcsik and Murugesan proposed a scheme of four categories for classifying citations (Moravcsik & Murugesan 1975). Chubin and Moitra developed a scheme of six categories for citation classification (Chubin & Moitra 1975). Nanba, Kando and Okumura developed a simplified classification scheme involving three categories (Nanba & Okumura 1999). Teufel, Siddharthan and Tidhar have proposed an annotation scheme of twelve categories for classifying citation sentences (Teufel et al. 2006).

Significant work has also been done for achieving automatic identification of contexts for sentences with citations. Garzone and Mercer developed a rule-based automated citation classifier (Garzone & Mercer 2000). Nanba, Kando and Okumura identified reference types using cue words (Nanba et al. 2000). Mercer and Marco analyzed the importance of fine-grained cue phrases in citation sentences for identifying citation reasons (Mercer & Marco 2003). Pham and Hoffmann used rules involving cue phrases for classifying scientific citations (Pham & Hoffmann 2003). Teufel et al. used machine learning techniques for achieving automatic classification of citation function (Teufel et al. 2006). Le et al. used finite-state machines for detecting citation types (Le et al. 2006). Kaplan et al. employed coreference chains for extracting citations from research papers (Kaplan et al. 2009).

2.1 New Context Types – The Rationale

Though there are different annotation schemes available, it is difficult to use them for a specific application. Baldi observes that most of these typologies were designed in an ad hoc manner, virtually isolated from one another. Further, these schemes were defined for different scenarios with different objectives (Baldi 1998). Further Guo et al. note that the scheme should be task oriented and the schema should cover the details required by the concerned application (Guo et al. 2010).

The focus of our study is to develop a tool that provides citation context based information services for the research community. Such a tool would provide information about the contexts of citation sentences and the adjacent sentences to it. Thus, the schema required by would include both citation and non-citation sentences. However, most of the studies have considered these sentences separately. Thus, we resort to defining a new set of context type definitions for using it to provide citation context based information services.

2.2 Use of CRFs – The Rationale

In recent times, conditional random fields (CRFs) have been successfully employed for achieving sequential classification of sentences. Hirohata et al. used CRFs for identifying sections in abstracts and achieved higher accuracy compared to Support Vector Machines (SVMs) (Hirohata et al. 1990). Chung (2009) and Kit et al. (2011) have successfully used CRFs for classifying sentences in medical abstracts (Chung 2009; Kim et al. 2011).

We also view sentence classification as a sequential classification problem, where the task is to predict a label sequence Yi, given an input sequence Xi. This implies that the label for a given sentence not only depends on the features of an individual sentence but also on the adjacent labels. This also facilitates capturing the rhetorical relations between sentences. Thus, we use conditional random fields (CRFs) for our experiments.

3 Sentence Context Types

In order to define context types for sentences with citations and their surrounding sentences, we chose randomly 20 articles from four different Lecture Notes in Computer Science (LNCS) volumes published by Springer (Springerlink.com, 2011). LNCS papers provided a uniform reference style (numbering style) which facilitated easy identification of sentences with citations. This formed our initial training dataset and studied paragraphs with citations in them. The training dataset of 20 articles provided 246 paragraphs.
We viewed each of these paragraphs as a set of reference areas, where a reference area is defined as the text that contains a citation sentence and its surrounding sentences (Nanba & Okumura 1999) and manually identified the possible contexts that were associated with sentences. This analysis resulted in defining the following set of context types for sentences.

### 3.1 Context Types for Non-Citation Sentences

The following contexts were defined for sentences without citations in research articles

**Background (BGR)**
- the sentence provides a background to an issue

**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

### 3.2 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 following contexts for citation sentences.

**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

**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 Works of Cited Work (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 subject area discussed in the paper

### 3.3 Framework for Modelling Sentence Context Types

The framework showing different patterns of relations between the sentence context types defined above in reference areas in a paragraph is shown in Figure 1. The arrows in Figure 1 indicate possible links between adjacent sentences. Several patterns can exist in a single paragraph.

For example, a typical pattern in reference areas that contain most of the context types defined above is indicated by label○. Such patterns appear in the beginning of the paragraph and the passage starts with sentences that provide a background or refer to an issue or gaps in research topics addressed in the article and are usually used for setting the context. These sentences may be followed by sentences that further describe the background or the issue or the gaps identified and sometimes can also refer to the results and future work of the current paper. After establishing the context, sentences with citations may appear. As seen earlier in Related Work section, citations are used for various purposes and we define a set of context types for citation sentences based on the reasons for using them in the article. As seen in Figure 1, the context types for citation sentences is modeled as an hierarchy which consists of using a citation for any issue (IRCW) at the root level. This branches out into various context types that indicate specific reasons for using the citation. Thus, citation sentences can have any of these contexts associated with them. These sentences can be followed by sentences without citations that can further describe or point out an issue or gaps in the citation cited earlier. They can also refer to the results or future work in relation to the work cited earlier. Various other patterns of reference areas are shown in Figure 1.

![Figure 1: Patterns of Contexts in Reference areas in Paragraphs with Citations in Research Articles](image-url)
The probability of the state variables drawn from the full model. This graphical structure also facilitates a functional form for their "clique functions" \( \psi_t \):

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

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.

## 4 Experiments with CRFs for Context Type Identification

As mentioned in Section 2.2 we used Conditional Random Fields (CRFs) for achieving the task of automatic context type identification. The objective was to learn probabilities of sentence labels for known features that encode properties of sentences that could be used for generating maximum likelihood labels for unseen sentences. The following explains the experiments carried out with CRFs.

### 4.1 Conditional Random Fields (CRFs)

CRFs were introduced for overcoming the label bias problems observed in MEMMs (Lafferty et al. 2001). 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 also 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 given the observations in a CRF with a linear chain structure (where the probability of each state \( y_t \) depends only on the probability of the state \( y_{t-1} \) and \( y_{t+1} \), and the observed data sequence) factors according to the following equation (Vail, 2008)

\[
P(Y|X) = \frac{1}{Z_x} \prod_{t} \psi_t(y_{t-1}, y_{t}, X)
\]

Here, \( X \) is the observation sequence, \( Y \) is the label sequence, and \( \psi_t \) 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:

\[
Z_x = \sum_{Y} \prod_{t} \psi_t(y_{t-1}, y_{t}, X)
\]

Conditional Random Fields use a particular functional form for their "clique functions" \( \psi_t \):

### 4.2 Dataset and Experimental Study, Training and Evaluation

#### 4.2.1 Dataset and Experimental Setup

We conducted experiments with 1000 paragraphs with citations extracted from 70 research articles chosen from LNCS. We used the first 40 articles as our training dataset and tested the model on the remaining 30 articles. Within the training set of 40 articles, we used the first set of 20 articles as our development dataset and used the complete set of 40 articles for evaluating and refining our classifier model. The following steps were followed sequentially in our experiments.

1. Feature Definition – to start with, we identified features based on the annotation scheme described in Section 3 in the first 20 articles and manually labelled them into one of these classes.
2. Feature Selection – we performed experiments with this initial dataset for analysing feature selection.
3. Developing the Classifier Model – after identifying the optimal feature set, we defined features and manually tagged sentences in the next 20 articles, resulting in the training set of 40 articles, which was used to develop our classifier model.
4. Testing – evaluated the classifier model on the test dataset of 30 articles.

#### 4.2.2 Training

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.

#### 4.2.3 Evaluation

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

\[
P = \frac{|TP|}{|TP|+|FP|}; \quad R = \frac{|TP|}{|TP|+|FN|}; \quad F = \frac{2PR}{P+R}
\]

where \( P \) represents precision, \( R \) represents recall, \( TP \) is the set of true positive, \( TN \) is the set of true negatives, and \( FP \) is the set of false negatives. A 10-fold 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.

#### 4.3 Feature Definition

In order to define features for sentences we used the initial set of 20 articles as our development dataset and examined the sentences in paragraphs with citations in these articles. The development dataset provided a total of 246 paragraphs. We manually examined each of these paragraphs. The citation sentence typically has preceding and following non-citation sentences with different contexts as indicated in the diagram. There are certain paragraphs which have only citation sentences (indicated by label \( \mathcal{Q} \)). The paragraphs can also start with a citation sentence followed by sentences without citations as indicated by label \( \mathcal{O} \). Thus different patterns of reference areas can appear in a paragraph with citations in research articles.

It needs to be noted that it is not necessary that a paragraph with citations would typically consist of only these patterns. There could be other patterns in these paragraphs and this depends on the writing style of authors. The patterns explained here are some of the basic patterns observed in the training dataset and the CRF-based tool used in this study is expected to learn other patterns.

---

**CRPIT Volume 144 - The Web 2013**
sentences and identified the following types of features for sentences.

4.3.1 Citation Features
Citation features indicate whether a given sentence has citation or not.

4.3.2 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 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.

4.3.3 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 as shown in Table 1. Table 1 lists the 9 categories of terms and phrases identified under term features that signify the context of the sentence. Table 1 also provides description of each of these categories along with the number of terms with examples identified in each of them.

4.3.4 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 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 using regular expressions for identifying and creating features for sentences in the text.

A feature set defined for a sample paragraph is shown in Listing 1. The terms in the paragraph which are responsible for defining the features are highlighted. As seen, sentence 1 does not have any of the terms and only the section feature is added to the paragraph. The paragraph belongs to the related work section of the article.

<table>
<thead>
<tr>
<th>Category and Description</th>
<th>E.g. Terms</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connecting Terms (CT) – Terms or phrases that indicate relations between sentences.</td>
<td>They, These, This, The authors.</td>
<td>27</td>
</tr>
<tr>
<td>Shortcoming Terms (SCT) – Terms or phrases that describe the shortcomings or gaps.</td>
<td>Nevertheless, performance suffers.</td>
<td>114</td>
</tr>
<tr>
<td>Methodology Terms (MET) – Terms or phrases that describe the methodology adopted or followed in the paper.</td>
<td>we consider, we use, we devise, we select</td>
<td>75</td>
</tr>
<tr>
<td>Result Terms (RES) – Terms or phrases that describe the results achieved either by the current paper or the cited paper.</td>
<td>we will show, we discover, we propose, we present</td>
<td>67</td>
</tr>
<tr>
<td>Future Work Term (FWT) – Terms or phrases that describe the future work of the paper</td>
<td>future work, we plan to extend.</td>
<td>27</td>
</tr>
<tr>
<td>Overcoming Gap Terms (OGT) – Terms or phrases that describe the characteristic of overcoming the identified gaps or shortcomings</td>
<td>enhanced, superior, outperforms, improves</td>
<td>33</td>
</tr>
<tr>
<td>Identifier Terms (IDT) – Terms or phrases that identify gaps or shortcomings in the related work or the cited work.</td>
<td>as shown, observations in,</td>
<td>30</td>
</tr>
<tr>
<td>Extending Terms (EXT) – Terms or phrases that discuss extending the current work with cited work.</td>
<td>builds on previous work,</td>
<td>4</td>
</tr>
<tr>
<td>Comparing Terms (COM) - Terms or phrases that mention comparison studies.</td>
<td>compared, comparison, in contrast to</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total number of terms</strong></td>
<td><strong>387</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Term Features defined for Development Set

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentHasCitation</td>
<td>Sentence has citation</td>
</tr>
<tr>
<td>prevSentHasCitation</td>
<td>Previous sentence has citation</td>
</tr>
<tr>
<td>sentSec=Intro</td>
<td>Sentence belong to the Introduction, Related Work, Motivation, Background, Our Approach sections of the article</td>
</tr>
<tr>
<td>sentSec=Sub</td>
<td>Sentence belongs to the Body block</td>
</tr>
<tr>
<td>sentSec=Conc</td>
<td>Sentence belongs to Conclusion block</td>
</tr>
<tr>
<td>sentHasTerm=CT</td>
<td>Sentence contains a connecting term or phrase</td>
</tr>
<tr>
<td>sentHasTerm=SCT</td>
<td>Sentence contains a shortcoming term or phrase</td>
</tr>
<tr>
<td>sentHasTerm=MET</td>
<td>Sentence contains a methodology term or phrase</td>
</tr>
<tr>
<td>sentHasTerm=RES</td>
<td>Sentence contains a resulting term or phrase</td>
</tr>
<tr>
<td>sentHasTerm=FWT</td>
<td>Sentence contains a future work term or phrase</td>
</tr>
<tr>
<td>sentHasTerm=OGT</td>
<td>Sentence contains an overcoming gap term</td>
</tr>
<tr>
<td>sentHasTerm=IDT</td>
<td>Sentence contains an identifier term</td>
</tr>
<tr>
<td>sentHasTerm=EXT</td>
<td>Sentence contains an extending term</td>
</tr>
<tr>
<td>sentHasTerm=COM</td>
<td>Sentence contains a comparing term</td>
</tr>
</tbody>
</table>

Table 2: Features defined for Sentences
Sample Paragraph
Several researchers have been studying the behavior of chains in MHWNs. Li et al. examine the performance of chains as the number of hops are increased and study the effect of cross-interference between chains [5]. They analyze the effect of MAC 802.11 behaviour on the performance of multi-hop chains but do not categorize interference patterns that govern network performance in terms of throughput and bandwidth utilization. Pang et al. present a hop by hop analysis of a multi-hop chain, study the impact of hidden terminals on the throughput chains, and present a quantitative approach towards estimating this throughput [6]. They show that hidden terminals cause packet drops affecting chain throughput and causing route stability.

Source: Majeed et al. (2009) [22]

Features defined for Sentences in the Sample Paragraph
Sentence 1: sentSec=Intro sentHasCitation IRCW
Sentence 2: sentSec=Intro sentHasCitation IRCW prevSentHasCitation IRCW
Sentence 3: sentSec=Intro sentHasTerm=CT sentHasTerm=SCT
Sentence 4: sentSec=Intro sentHasTerm=CT sentHasCitation IRCW
Sentence 5: sentSec=Intro sentHasTerm=CT sentHasTerm=ISS prevSentHasCitation DES

Listing 1: Feature set defined for a sample paragraph

4.4 Feature Selection
In order to examine the performance of the classifier for the features identified above, we conducted classification experiments with the development set. To this end, the developments set was divided into a training dataset and development-test (dev-test) dataset 200 and 46 paragraphs respectively. The classification results for the development set are provided in Table 3.

<table>
<thead>
<tr>
<th>Label</th>
<th>Freq (%)</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Results</td>
<td>Results with Issue Features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall Accuracy: 78%</td>
<td>Overall Accuracy: 86%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>BGR 5</td>
<td>0.57</td>
<td>0.88</td>
<td>0.69</td>
</tr>
<tr>
<td>CCW 1</td>
<td>1.00</td>
<td>0.50</td>
<td>0.66</td>
</tr>
<tr>
<td>CWECW 0.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>CWIG 0.5</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>CWO 6.7</td>
<td>1.00</td>
<td>0.75</td>
<td>0.85</td>
</tr>
<tr>
<td>DES 35.1</td>
<td>0.73</td>
<td>0.79</td>
<td>0.76</td>
</tr>
<tr>
<td>FW 1.6</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>IRCW 12.8</td>
<td>0.80</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>ISSUE 8.9</td>
<td>0.11</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>RWSC 19.5</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>SCCW 0.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>WUCW 7.2</td>
<td>1.00</td>
<td>0.92</td>
<td>0.96</td>
</tr>
</tbody>
</table>

P = Precision; R = Recall; F = F-Score; Freq = Frequency

Table 3: Classification Results for Development Dataset
The initial set of results (indicated by A in Table 3) shows that the classifier achieved an overall accuracy of 78% and the classes that performed poorly included the citation sentence classes of CWECW, CWIG and SCCW and the non-citation sentence classes of ISSUE (shown in grey in Table 3).

One of the reasons for this poor performance can be attributed to the lower number of instances. The citation sentence classes appeared only 0.5% of the time and this accuracy can be increased with the increase in training data. However the non-citation class of ISSUE with a 9% frequency achieved an F-Score of 0.08. The confusion matrix for this data (not provided in this paper) showed that the ISSUE sentences were classified most of the time as Description (DES) sentences. Thus, in order to differentiate between ISSUE and DES classes, we defined another category of terms called ‘ISSUE Terms’ for defining ISSUE features. We identified 34 terms belonging to this category with terms such as Therefore, Thus, Hence, for this purpose. The inclusion of these features increased the accuracy for the ISSUE class from an F-Score of 0.08 to 0.76. Correspondingly, the overall accuracy of the classifier also increased to 86%. The results of the classifier with the additional feature are shown in Table 3 (indicated by B).

4.5 Developing Classifier Model
After identifying the optimal set of features, we considered the set of 40 articles as our dataset and identified features for sentences in them. The training set of 40 articles provided a total of 2539 sentences, which had 691 and 1848 sentences with and without citations, respectively. While experimenting with CRFs, we used two CRF structures: a first-order linear chain and a linear chain with additional zero-order features. Our earlier experiments with sentences in related work sections showed that a first-order linear chain performs poorly for classes with a low number of instances (Angrosh et al. 2010). Thus, in order to increase accuracy for classes which appear less, we use a linear chain CRF with additional zero-order features to provide a “back-off” prediction capability.

The classification results of the classifier for using first-order linear chain CRF and a CRF with additional zero-order features are provided in Table 4 and their performance is graphically shown in Figure 2.

Table 4: Results of Classifiers for Training Dataset
The classification results show that the first-order linear chain CRF with additional zero-order features achieves a slightly higher accuracy of 91.80% compared to an accuracy of 90.23% achieved by a first-order linear chain.
A significant increase is seen for classes of CCW, FW and RWRW all of which belong to the category of classes with lower accuracy. The classifier performed extremely well for classes DES, ISSUE, IRCW and RWSC.

The classifier results of the model on the test dataset are tabulated in Table 5. As seen, the classifier achieved an accuracy of 92.08%, 92.92% and 90.01% with test dataset 1, 2 and 3 respectively. The average accuracy achieved over the test dataset of 30 articles was 91.67%.

5 CitContExt – Citation Context Extractor

CitContExt is a web application that integrates a citation context extraction tool with SciVerse Content APIs (Sciverse n.d.) for identifying citation contexts in research articles. The SciVerse Content APIs provide direct access to Elsevier content using RESTful URL requests. The applications developed using SciVerse Content APIs can be integrated in either ScienceDirect or Scopus interface.

![Figure 2: Performance of Classifiers for Training Dataset](image)

**Table 5: Classifier Results on Test Dataset**

<table>
<thead>
<tr>
<th>Label</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGR</td>
<td>1.00</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.82</td>
<td>0.84</td>
<td>0.80</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>CCW</td>
<td>0.66</td>
<td>1.00</td>
<td>0.80</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.50</td>
<td>0.66</td>
</tr>
<tr>
<td>CWIG</td>
<td>0.77</td>
<td>0.70</td>
<td>0.73</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.83</td>
<td>0.55</td>
<td>0.66</td>
</tr>
<tr>
<td>CWO</td>
<td>0.82</td>
<td>0.85</td>
<td>0.83</td>
<td>0.91</td>
<td>0.84</td>
<td>0.88</td>
<td>0.90</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>CWOG</td>
<td>0.75</td>
<td>1.00</td>
<td>0.85</td>
<td>0.93</td>
<td>0.82</td>
<td>0.87</td>
<td>0.64</td>
<td>0.57</td>
<td>0.61</td>
</tr>
<tr>
<td>DES</td>
<td>0.92</td>
<td>0.99</td>
<td>0.95</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
<td>0.94</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>FW</td>
<td>1.00</td>
<td>0.66</td>
<td>0.80</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>IRCW</td>
<td>0.94</td>
<td>0.96</td>
<td>0.95</td>
<td>0.91</td>
<td>0.97</td>
<td>0.94</td>
<td>0.90</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>ISSUE</td>
<td>0.91</td>
<td>0.98</td>
<td>0.95</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
<td>0.91</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>RWRW</td>
<td>1.00</td>
<td>0.37</td>
<td>0.54</td>
<td>1.00</td>
<td>0.33</td>
<td>0.50</td>
<td>0.57</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>RWSC</td>
<td>1.00</td>
<td>0.82</td>
<td>0.90</td>
<td>0.98</td>
<td>0.91</td>
<td>0.95</td>
<td>0.94</td>
<td>0.88</td>
<td>0.91</td>
</tr>
<tr>
<td>SCCW</td>
<td>0.83</td>
<td>0.55</td>
<td>0.66</td>
<td>0.88</td>
<td>0.78</td>
<td>0.83</td>
<td>0.61</td>
<td>0.69</td>
<td>0.65</td>
</tr>
<tr>
<td>WUCW</td>
<td>0.81</td>
<td>0.60</td>
<td>0.69</td>
<td>0.80</td>
<td>0.72</td>
<td>0.76</td>
<td>1.00</td>
<td>0.69</td>
<td>0.81</td>
</tr>
</tbody>
</table>

P – Precision; R – Recall; F – F-Score

5.1 Architecture of CitContExt

The architecture of CitContExt is shown in Figure 3.
The SciVerse Application Framework (Sciverse n.d.), which extends the default Shindig Framework (Shindig n.d.) offers a customized environment for developing JavaScript applications (also known as Gadgets or Widgets) and can be hosted on SciVerse web sites as extensions of the already available SciVerse services. Besides using data from SciVerse Content APIs, the application can be configured to interact with an external application to send/receive external data. As seen in Figure 3, CitContExt consists of a JavaScript module hosted on SciVerse platform for accessing SciVerse content and a locally maintained Web Service that interact with the JavaScript module for identifying contexts of sentences in the accessed content.

The following explains the key functions of each of these components.

### 5.1.1 Gadget / JavaScript Module
1. Use SciVerse Content API to obtain the metadata and the full-text of the source article and the citing articles.
2. Extract paragraphs and references – The XML data returned from content API is processed to identify paragraphs with citations and references. Information about the section to which paragraphs belong is also obtained. While all paragraphs with citations are extracted from the source article, the citing articles are processed and only those paragraphs where a reference to the source article as a cited work is found are identified and extracted.
3. Send content to web service – The extracted content is sent to the web service as a JSON object for identifying the context-types of sentences.
4. Handle response received from the web service – The application receives contextual information as a JSON object
5. Display contextual information – The received JSON object is processed and displayed in SciVerse websites. The application uses treeplugin and Pie chart plugin for displaying the data as a tree structure and pie chart respectively.

### 5.1.2 Web Service
The web service is a locally managed Java application that receives content from the JavaScript module and returns contextual information. The important components of the application are as follows:
1. The application runs on a Tomcat Server and use HTTP GET for receiving and returning content as a JSON object.
2. Segment sentences – The paragraphs received by from the Gadget needs to be segmented before context types for each sentence can be identified. OpenNLP (OpenNLP n.d.), a Java-based Natural Language Processing (NLP) tool is used for achieving sentence segmentation.
3. Generate features – Feature generation is an important step in the process of identifying context-types. Before being provided as an input to the classifier, each of the segmented sentences is converted into a set of features that describe the properties of a sentence.
4. Identify context type – The obtained features are submitted as an input to the pre-trained sentence context classifier model for identifying context types for each sentence.

### 5.2 Information Services of CitContExt
CitContExt offers the following citation context based information services for the research community. A screenshot of the application is provided in Figure 4. The following are the information services offered by CitContExt.

This figure shows the identification of contexts for sentences (in a paragraph) in an article published in ScienceDirect. The first and the third sentences are identified as ISSUE sentences, and the sentence is labelled as CWOG – which implies that the cited work overcomes the gaps described.

**Figure 4: Context Identification of Sentences in ScienceDirect Articles**
5.2.1 View Contexts of Citation and Non-Citation Sentences
The application enables users to view the contexts of citation and non-citation sentences in the article. This helps the user to understand the context in which a citation is used. Further, the contexts of the surrounding sentences of a citation sentence helps in further understanding the context.

5.2.2 View Contexts of Citing Articles
Unlike current applications, which simply provide information about the number of citing articles, CitContExt goes further and extracts the text from the citing article, citing a given article and presents to the user the context in which it is cited. This helps in eliminating the ardent task of referring to numerous articles in order to learn about the use of citation across articles.

5.2.3 Browse Classification of Citation Sentences
The application also allows the user to view a classification of sentences according to citations types in a paragraph. This helps in quickly identifying the different types of citations used in a given paragraph in the article.

5.2.4 Browse Classification of Citation Sentences
The application further allows the user to view a classification of citing articles according to citation types. This again helps the user in selecting the citing article based on the citation type.

6 Use Cases
Our application finds the following use cases.

6.1 Useful Tool for Research Students
The tool would immensely benefit research students. This tool provides information about the use of cited works in an article, which can help identify cited works used for preparing the background of the article or cited works which identify gaps or cited works which themselves have gaps. This information would be of immense value for understanding the topic. Further, the information about the contexts of citing articles would greatly ease the burden of reading all citing articles and facilitates better understanding of an article.

6.2 Help Scholars New to a Topic
Frequently, there are times when scholars want to quickly learn about a new topic. This tool can facilitate in providing a quick overview of the topic by displaying the contextual relationships.

6.3 Identify Citation Relationships
The tool addresses the crucial problem of identifying citation relationships that can be used in several applications. This includes (a) creating automatic literature review (b) text summarization (c) semantic annotation (4) sketching intellectual lineage for ideas (c) citation context analysis etc.

6.4 Easy Integration with Current Services
The use of Web APIs facilitated in Sciverse providing services without the need for knowing about user’s subscription details for Elsevier content. The Web APIs also facilitate in providing the services at runtime, overcoming scalability problems. Thus, the application can be easily integrated with the available services.

7 Conclusion
We presented in this paper our research work carried out for identifying contexts associated with sentences in research articles and use the identified knowledge to provide intelligent information services for the research community. In order to achieve our objective, we developed an annotation scheme that focused on defining various contexts for sentences in scientific articles. In order to establish the reliability of the annotation scheme, we have carried out an inter-rater study. The study resulted in achieving an overall agreement of 89.93% among annotators (Angrosh et al. 2012a).

We also employed supervised learning methods using sequential learning models for achieving automatic sentence classification. We have compared the results of our classifier with similar studies such as those of Teufel and Moens (1999) and Teufel et al. (2006). We note that our results achieve higher accuracy in comparison to these studies (Angrosh et al. 2012b).

We also described CitContExt – a Citation Context Extracting tool developed for Sciverse application platform based on the above techniques and examined its application through various use cases.

Our future work involves studying the application of our annotation scheme and our tool across other disciplines such as humanities and medical sciences.

With respect to CitContExt, we intend to implement an interface that enables users to provide feedback about the classification of sentences and information about the terms and phrases that contribute in the setting context-types in sentences. This could help in improving our current system.

8 References


