A Citation Centric Annotation Scheme for Scientific Articles

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

This paper presents an annotation scheme for modelling citation contexts in scientific articles. We present an argumentation framework based on the Toulmin model for scientific articles and develop an annotation scheme with different context types based on the argumentation model. We present the results of the inter-rater reliability study carried out for studying the reliability of our annotation scheme.

1 Introduction

Citations play an important role in scientific writing. However, there are not many tools that provide citation context based information services. The citation services provided by academic search engines such as Google Scholar[1] includes information about the number of citing documents and links to citing articles. Search can also be performed for keywords in citing articles. Citation focused tools such as CiteSeerX [2] and Microsoft Academic Search [3] engines go a little further in identifying the passage of citations in citing articles. The objective of such services is to facilitate quick access to citation content to aid the learning process. However, the high volume of research content renders it difficult to achieve optimum use of these services.

Identifying this need, we proposed to develop tools for providing intelligent citation context based information services. However, an annotation scheme for citation contexts is one of the key requirements of citation context based information tools. An annotation scheme providing citation contexts can help in classifying citation passages and provide better citation context based information services. Accordingly, we studied the existing annotation schemes and noted that it was difficult to use these schemes for our application and proposed to develop a new annotation scheme. Angrosh, Cranefield and Stanger (2012a) have successfully used the annotation scheme with machine learning techniques for developing intelligent citation context based tools. These included a linked data application (Angrosh, Cranefield, and Stanger, 2012b) and a Web-based application[4].

We present in this paper our annotation scheme designed to represent citation contexts based on an argumentation model, which can be used to develop citation context based information tools.

2 Related Work

Over the years, several researchers have proposed annotation schemes for scientific articles. These schemes can be classified into two categories: (a) those that consider the full text of an article; and (b) those addressing citation sentences only.

2.1 Annotation Schemes for Full Text

Conceptualizing the idea of 'argumentative zoning', Teufel (1999) proposed an annotation scheme of seven categories and called them argumentative zones for sentences. Mizuta and Collier (2004a) extended Teufel's argumentation scheme (Teufel, 1999) for zone analysis in biology texts and provided a scheme of seven categories. Langer et al. (2004) noted that newer applications in areas such as the Semantic Web required richer and more fine-grained annotation of seven topic types for documents. Motivated by the need to identify passages of reliable scientific facts, Wilbur et al. (2006) devised an annotation scheme of five categories for biomedical texts. Ibekwe-sanjuan et al. (2007) developed local grammars to annotate sentences in a rhetorical scheme consisting of eight categories. Liakata et al. (2009) presented two complementary annotation schemes for scientific papers in

[1] http://scholar.google..com
[2] http://citeseerx.ist.psu.edu
[3] http://academic.research.microsoft.com/
[4] www.applications.sciverse.com/action/appDetail/297884?zone=main&pageOrigin=appGallery&activity=display

Angrosh M.A., Stephen Cranefield and Nigel Stanger. 2012. A Citation Centric Annotation Scheme for Scientific Articles. In *Proceedings of Australasian Language Technology Association Workshop*, pages 5−14.

Chemistry: the Core Scientific Concepts (CoreSC) annotation scheme and the Argumentative Zoning-II scheme (AZ-II) (Teufel, 1999).

2.2 Annotation Schemes for Citation Sentences

Researchers have also specifically focused on citation sentences. In 1965, Eugene Garfield, the creator of Science Citation Index, outlined fifteen different reasons for citations (Garfield, 1964). Lipetz (1965) explored the possibility of improving selectivity in citation indexes by including citation relationship indicators and devised a scheme of 29 citation relationship types for science literature. Claimed to be the first in-depth study on classifying citations, Moravcsik and Murugesan (1975) proposed a classification scheme for citations consisting of four categories. Chubin and Moitra (1975) redefined the four categories of Moravcsik and Murugesan as a set of six mutually exclusive categories in order to further generalize the scheme. Spiegel-Rosing (1977) analyzed the use of references in 66 articles published in *Science Studies* and proposed a classification scheme of 13 categories. Oppenheim and Renn (1978) proposed a scheme of seven categories identifying citation reasons for historical papers. Frost (1979) proposed a classification scheme of citations in literary research and applied the scheme for a sample of publications in German literary research. Peritz (1983) proposed a classification scheme of eight categories for substantive-empirical papers in social sciences.

Focusing on automatic citation identification and classification, Nanba and Okumura (1999) proposed a simplified classification scheme of three categories based on the 15 reasons identified by Garfield (1964). Pham and Hoffmann (2003) developed KAFTAN, a "Knowledge Acquisition Framework for TAsks in Natural language", which classified citations into four citation types. Teufel, Siddharthan, and Tidhar (2006) presented an annotation scheme for citations involving 12 categories, based on the categories proposed by Spiegel-Rosing (1977). Le et al. (2006) defined six categories of citations that facilitated emerging trend detection. Radoulov (2008) carried out a study for exploring automatic citation function classification and redesigned Garzone's scheme of 35 categories from the perspective of usability and usefulness.

3 Why another Annotation Scheme?

Though there already exist various annotation schemes for scientific articles, we present in this paper another annotation scheme that defines various context types for sentences. The objective of developing this annotation scheme is to provide a set of context type definitions that can be used for providing citation context based information services.

There exist several difficulties in using existing schemes across different applications. Baldi (1998) notes the older classification schemes published during the 1970s and the 1980s were developed in a completely ad hoc manner and were virtually isolated from one another during the development process. White (2004) described existing classification schemes as "idiosyncratic" and emphasized the difficulty in employing them, particularly when using them across disciplines.

Studies that have focused on the full text of the article (Teufel, 1999; Mizuta and Collier, 2004a; Langer et al., 2004; Wilbur et al., 2006; Ibekwe-sanjuan et al., 2007) have proposed a generic set of categories that would be less useful in designing citation context based services. The objective of these studies has been to achieve text summarization.

On the other hand, studies carried out with citation sentences have proposed fine-grained categories that are difficult to use. The use of these classification schemes presents challenges in defining features in order to achieve automatic citation classification. Further, a focus on citation sentences alone would result in excluding surrounding sentences of citations that can provide additional contextual knowledge about citations.

Gao, Tang and Lin (2009) recommend that the selection of an annotation scheme should be task-oriented, and that the optimal scheme for use should depend on the level of detail required by the application at hand. The key focus of our study is to identify contexts of citations and develop information systems based on this contextual knowledge. However, the use of existing schemes creates difficulties as it is either too generic or fine grained as mentioned above. Our application would require an annotation scheme that would consider both citation and non-citation sentences and provide a set of context types that can also be used for automatic context identification.

The structure of this paper is as follows. In Section 4, we describe an argumentation framework for scientific articles based on the Toulmin model. In Section 5, we present the different context type definitions for sentences in scientific articles, defined based on the argumentation model. In Section 6, we discuss the results of an inter-rater reliability study to evaluate this set of context types, and we conclude the paper in Section 7.

4 Argumentation in Scientific Articles

An important characteristic of a scientific article is its persuasive nature, and citations play an important role in developing this feature for an article. Gilbert (1977) viewed scientific papers as 'tools of persuasion' and noted that references increase the persuasiveness of a scientific paper. Brooks (1985) surveyed authors to assess their motivations for citing and concluded that persuasiveness was a major motivating factor. In another study, Brooks (1986) further confirmed his findings with a different survey, concluding that persuasiveness was the dominant reason for citing. Cozzens (1989) observed that the primary function of a document is to persuasively argue about a knowledge claim, and noted that the art of writing scientific papers consists of marshalling the available rhetorical resources such as citations to achieve this goal. Hyland (2002) observed that arguments in research articles required procedural and citation support.

Thus, it is evident that scientific papers are argumentative in nature and one of the prominent reasons for using citations is to persuade the reader about the argument presented in the paper.

4.1 Toulmin Model for Modelling Scientific Discourse

In order to develop an argumentation model for scientific articles, we make use of the well-known Toulmin model of argument (Toulmin, 1958). The Toulmin model asserts that most arguments can be modelled using six elements: claim, data, warrant, qualifier, rebuttal and backing. The first three are considered as essential components and the last three as additional components of an argument.

The Toulmin model of argument can be applied for scientific articles as shown in Figure 1. The different elements of the Toulmin model as applied to scientific articles are explained below.

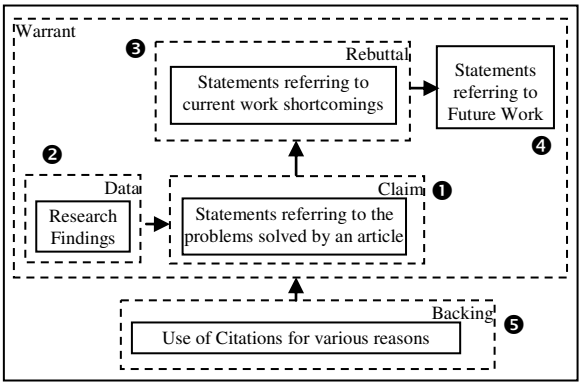

Figure 1: Argumentation Framework for Scientific Articles

A **Claim** in the Toulmin model refers to the proposition or conclusion of an argument. With respect to a scientific article, a claim therefore consists of the statements that describe the outcomes or the results of the article (block ❶ in Figure 1). In other words, these are the statements that refer to the problems solved by the article.

The **Data** in the Toulmin model refers to the factual information that supports the claim. In a scientific article, these are the research findings that are used to substantiate the claim, i.e., the results or outcomes of the article (block ❷).

A **Rebuttal** in the Toulmin model refers to exceptions to the claim. In a scientific article, these are the statements that refer to the shortcomings or gaps of the article (block ❸). These are situations where the results of the article may not hold or the problems that the article has not solved. The rebuttal statements also result in statements that refer to future work arising from the article (block ❹).

A **Warrant** in the Toulmin model forms the key of the argument process and provides the reasoning and the thinking process that binds the data to the claim. With respect to a scientific article, warrants play a crucial role as it is this reasoning aspect that is responsible for making the article presentable. These are specifically a set of sentences that relate the data and claim for making the article convincing. They also connect rebuttal sentences for making the argument clear. The dotted lines in Figure 1, surrounding blocks ❶ to ❹, indicates this aspect of warrants, bringing together the different components of the article.

A **Backing** in the Toulmin model refers to aspects that provide additional support to a warrant. The use of citations in scientific articles

can be identified with *backing* as explained below.

4.2 Role of Citations in the Argumentation Model

We explained in the previous sections an argumentation model for scientific articles based on the Toulmin model. We also noted that citations play an important role in scientific writing with their persuading nature as one its prime characteristics.

In the argumentation model discussed above, citations can play various roles in different components. For example, citations can facilitate development of a good warrant. Citations can also be considered as warrants themselves as they also contribute to the reasoning process for linking data and the claim. Further, the data in the article can be developed using the outputs of the cited work.

To generalize this notion, citations can be considered as *backing* providing additional support to the different components of the argumentation model. This is indicated in Figure 1 with a link provided from block ❺ to the overall warrant block comprising different elements of the Toulmin model.

5 Context Types for Sentences based on Argumentation Model

We discussed in the previous section an argumentation framework for scientific articles based on the Toulmin model of argument and identified citations as a *backing* component for presenting the argumentation made in the paper. We also noted that one of the prominent reasons for using citations is to persuade the reader. Generally, the act of persuasion involves providing sufficient proof in order to convince the reader about a concept or an idea. In a scientific article, the use of citations to persuade the reader may focus on the following:

1. To demonstrate that others (cited work(s)) have identified a similar problem.
2. To demonstrate that others (in cited work(s)) have solved a similar problem.
3. To demonstrate how the current paper solves other problems (presented in cited paper(s))
4. To demonstrate the limitations or the shortcomings or the gaps of others (in cited work(s))
5. To compare works of others (in cited paper(s))
6. To compare the results of the current work with others (in cited work(s))

Thus, we analyzed citations against these persuading characteristics in order to examine the role of citations. To this end, we created a dataset of 1000 paragraphs that had sentences with citations. Paragraphs with citations were only considered, as the focus was to identify the contexts of sentences with citations. These paragraphs were obtained from 71 research articles. The articles were chosen from the Lecture Notes in Computer Science (LNCS) series published by Springer and accessed according to the terms of our institutional licence.

The dataset had a total of 4307 sentences, including 1274 citation sentences and 3031 non-citation sentences. We differentiated between citation and non-citation sentences and manually analyzed each of these sentences. This resulted in defining various context types for sentences as discussed below.

The Oxford English Dictionary 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" (Oxford Dictionary of English, 2010). Thus, the set of words in a sentence are chosen with an intention to provide a meaning to the reader. Following this definition, we identify the context type of a sentence as 'the meaning of a sentence'. For example consider the passage shown in Figure 2. Each of the sentences in the passage has its own meaning and thus a specific context type. For instance, in Sentence 1, the authors refer to a cited work to describe a topic and in Sentences 2 and 3, the authors further describe that cited work.

1 *Razak et al. have studied the effect of MAC interactions on single chains under saturated UDP traffic [1].*
2 *They develop a systematic methodology for determining the types of interaction that are possible in chains of 3 and 4 hops and they study the effect of these interactions on chain performance.*
3 *They further extend their work to analyze chains of n hops.*
4 *These studies do not consider the effect of TCP traffic on chain performance.*
5 *TCP introduces several factors like bi-directional traffic, congestion control, round trip time estimations for timeout prediction etc. that are affected by interference interactions within a chain.*
6 *As we will show in this paper, the types of interactions within chain have a substantial effect on the performance of a network under TCP traffic.*

Source: Majeed et al. (2009)

Figure 2: Example Passage

Further, it needs to be noted that the context of a citation may not be evident from the sentence containing the citation alone and may require

understanding of sentences surrounding this sentence, which necessitates the need to identify the contexts of surrounding sentences. For example, in the passage provided in Figure 2, though the authors refer to the cited work in sentence 1 and further describe it in sentences 2 and 3, it is only in sentence 4 (shaded in grey) that the authors identify gaps in the cited work(s). Thus, in order to understand the citation context of the citation in sentence 1, we need to identify the context types of surrounding sentences with and without citations.

5.1 Context Types for Citation Sentences

The analysis of citation sentences with a focus on the persuasive nature of citations described above resulted in identifying the following context types for citation sentences.

1. *Citing Works to Identify Gaps or Problems (CWIG)* – authors cite works that identify gaps to inform the reader about the existence of a problem.
2. *Citing Works that Overcome Gaps or Problems (CWOG)* – authors cite works to inform readers that other researchers are working on a similar problem and that they have solved some of the identified gaps.
3. *Using Outputs from Cited Works (UOCW)* – authors cite works to inform the reader about their outputs such as a methodology or training dataset, especially when these are used in the author's research.
4. *Comparing Cited Works (CCW)* – authors cite various related works and provide a comparison to bolster their argument.
5. *Results with Cited Works (RWCW)* – authors cite works to relate their research results to the cited work.
6. *Shortcomings in Cited Works (SCCW)* – authors cite works to identify shortcomings or gaps in them.
7. *Issue Related Cited Works (IRCW)* – authors cite works to inform readers about related research issues.

5.2 Context Types for Non-Citation Sentences

Similarly we analyzed the surrounding non-citation sentences and accordingly identified the following types of non-citation sentences:

1. *Background Sentences (BGR)* – Sentences that provide background or introduction.
2. *Gaps Sentences (GAPS)* – Sentences that identify gaps or problems. It was observed that authors identify gaps or problems in different ways. For example, authors identified gaps in the work cited earlier in the article, or related research topics addressed in the article, or could also mention the shortcomings of the current article itself.
3. *Issue Sentences (ISSUE)* – Sentences that refer to author viewpoints. These sentences are referred as issue sentences as these are the issues or points identified by the author.
4. *Current Work Outcome Sentences (CWO)* – Sentences that refer to the outcomes of the current research (the work being reported).
5. *Future Work Sentences (FW)* – Sentences that refer to future work.
6. *Descriptive Sentences (DES)* – Sentences that are descriptive in nature. For example, authors can further describe a cited work.

Thus, following this approach, we developed the annotation scheme shown in Figure 3 for sentences in full text of articles.

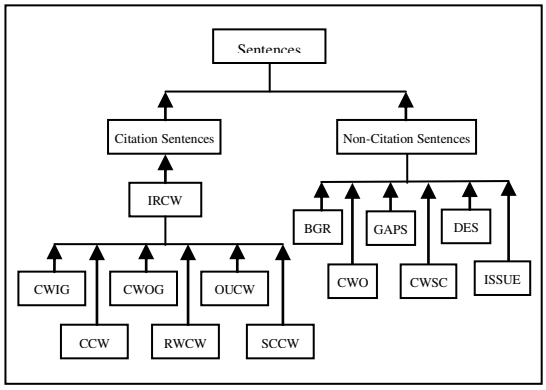

Figure 3: Our Initial Annotation Scheme

As seen in Figure 3, sentences are initially classified as citation sentences or non-citation sentences. With respect to non-citation sentences, we define six different context types that could be associated with them. However, with respect to citation sentences, we define a hierarchy of context types for sentence. We define at the top level of the hierarchy the class of IRCW (Issue Related Cited Work). Thus, if a citation sentence cannot be classified into any other class, it will have the class of IRCW. This implies that a citation in the article is made for some issue other than the context types defined in our annotation scheme. We identify six context types for citation sentences as subclasses of IRCW.

6 Reliability of Context Types

In order to study how reliably coders can interpret the context types defined above in an objective way, we carried out an inter-rater reliability (IRR) study. The approach followed during this study is as follows.

6.1 Approach of IRR Study

Researchers have adopted different strategies while carrying out inter-rater reliability studies. Teufel and Moens (2002) worked with annotations of a subset of 80 conference articles from a larger corpus of 260 articles. The articles were annotated by two annotators other than the first author herself. Wilbur et al. (2006) chose 10 articles randomly and worked with nine annotators for reporting annotation results.

With respect to the practice adopted for reporting inter-annotator agreement, Bird et al. (2009) note that double annotation of 10% of the corpus forms a good practice in such studies. Further Artstein and Poesio (2008) observe that the most common approach to infer the reliability of large-scale annotation involves each sentence being marked by one coder and measuring agreement using a smaller subset that is annotated by multiple coders. We adopted a similar approach for measuring the agreement about the context type definitions proposed in our study.

As mentioned earlier, the training dataset was created using 70 articles chosen from LNCS. We chose to annotate 10% of the corpus. Each article was annotated by at least three annotators, with one of the annotators being the first author of this paper. This facilitated deriving the following measures: (a) overall agreement between annotators (b) agreement between individual annotators, and (c) agreement for each label.

Choice of Statistic – We used Krippendorff's alpha (α) (Krippendorff, 2011) for measuring reliability as it provides a generalization of several known reliability indices. This statistic enables researchers to evaluate different kinds of data using the same reliability standard and can be used in different situations such as (a) any number of observers, (b) any number of categories, scale values or measures, (c) large and small sample sizes, and (d) incomplete or missing data. Krippendorff's alpha (α) is defined "as a reliability coefficient developed to measure the agreement among observers, coders, judges or raters" (Krippendorff, 2011). The general form of α is given by:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where D_o is the observed disagreement among values assigned to units of analysis:

$$D_o = \frac{1}{\mathrm{n}} \sum_{\mathrm{c}} \sum_{\mathrm{k}} o_{\mathrm{ck}} \delta^2_{\mathrm{ck}}$$

and D_e is the expected disagreement due to chance rather than properties of the units:

$$D_e = \frac{1}{\mathrm{n}(\mathrm{n}-1)} \sum_{\mathrm{c}} \sum_{\mathrm{k}} \mathrm{n_c}.\mathrm{n_k} \delta^2_{\mathrm{ck}}$$

The arguments of the two disagreement measures, o_{ck}, $\mathrm{n_c}$, $\mathrm{n_k}$ and n refer to frequencies of values in the coincidence matrices.

Characteristics of the annotators – The annotators in our study had considerable training in scientific writing and publishing with most of them being PhD students pursuing their doctoral research and a few of them being faculty members in the field of information science.

Choice of Articles – We provided annotatators articles chosen from their own field. This was done for the following reasons: (a) it would be easier for annotators to understand the content and hence apply the labels easily and thoughtfully; and (b) minimize the annotating time; and (c) as a motivation factor as articles were from their own field.

Guidelines for annotators – The guidelines used for annotators provided details of the context type definitions of the study along with example sentences for each definition. Each annotator was briefed about the purpose the study and the context type definitions. The briefing time spent with the annotators ranged between 15 and 30 minutes. The annotators were provided with paragraphs containing citations that were extracted from articles. The paragraphs were formatted to show individual sentences in them with the citation sentences highlighted to help annotators distinguish between citation and non-citation sentences.

Before carrying out the study, we conducted a pilot study to examine the feasibility of our approach. The pilot study resulted in making certain changes to the annotation scheme as will be discussed in the following section.

6.2 Pilot Study

We conducted a pilot study with three annotators using three articles, with each annotator annotating one article. All three articles were also annotated by the first author of this article,

henceforth referred to as Annotator A. Thus, we were able to compare the annotations made by Annotator A with the annotations of the three other annotators. The paragraphs extracted from the three articles provided a total of 300 sentences and hence there were 300 cases and 600 decisions to be made. After the coding of individual articles, we had discussions with the coders about the study. The coders felt that the context type definitions were clear enough and the examples were helpful in classifying sentences. However, they said there was confusion between the classes DES and ISSUE and, it was difficult to distinguish between the two. The analysis of these experiments resulted in a Krippendorff's Alpha (α) (Krippendorff, 2011), score of 0.79 (N = 300, k = 2), where N is the number of items (sentences) and k is the number of coders. This is equivalent to 85% agreement between the Annotator A and each of the three annotators. The classification results for each label along with the confusion matrix are shown in Table 1[5].

As can be seen in Table 1, there was confusion for the classes DES and ISSUE. With respect to Description (DES) sentences, the coders classified about 10% (14 out of 144) as ISSUE sentences and 62% (18 out of 29) of ISSUE sentences as DES sentences. Thus, in order to avoid this confusion, we merged the classes of DES and ISSUE into one class of DES and removed the label ISSUE from context type definitions. The merging of these classes resulted in achieving a α value of 0.93 for the pilot data, which is 95.7% agreement between the annotators. With these changes we carried out the study with a large number of annotators, as discussed in the next section. The modified annotation scheme based on the results of the pilot study is shown in Figure 4.

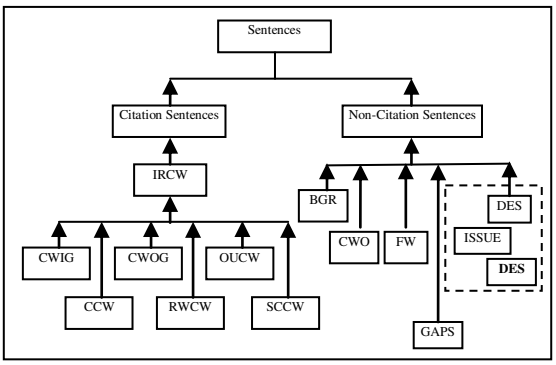

Figure 4: Modified Annotation Scheme

[5] Table 1 is provided at the end of this paper for formatting reasons.

6.3 IRR Study with Larger Sample

After conducting the pilot study and making necessary corrections to the context type definitions, we carried out a study using 11 annotators and 9 articles. This formed 12% of the training dataset. Each article was annotated by two annotators other than Annotator A, (the first author) who annotated all nine articles. The set of 9 articles provided a total of 907 sentences. The overall result achieved for α, involving nine articles and 11 annotators was 0.841 as shown in Table 2.

This is equivalent to 89.93% agreement between different pairs of annotators. The number of coders indicates that each article was annotated by three annotators. An agreement α = 0.8 or higher on Krippendorff's scale is considered as a reliable agreement, and an agreement of 0.67 to 0.8 is considered to be marginally reliable. A value lower than 0.67 for α indicates the agreement is unreliable. Therefore, the results indicate that the labels of our scheme can be reliably applied to sentences.

% Agreement	α	No. of Coders	No. of Cases	No. of Decisions
89.93	0.841	3	907	2721

Table 2: Overall Results

The details of the agreement between Annotator A and the others annotators involved in the study is shown in Table 3.

A	B	C	D	E	F	G
Annotator 4	85.3	0.760	93	16	109	218
Annotator 5	85.5	0.719	94	16	110	220
Annotator 9	87.8	0.836	72	10	82	164
Annotator 8	88.6	0.847	39	5	44	88
Annotator 10	90.5	0.831	95	10	105	210
Annotator 1	90.8	0.854	315	32	347	694
Annotator 6	91.8	0.832	101	9	110	220
Annotator 2	92.2	0.877	320	27	347	694
Annotator 7	94.4	0.930	119	7	126	252
Annotator 3	94.8	0.903	312	17	329	658
Annotator 11	95.2	0.907	100	5	105	210

A – Comparison between Annotator A and the Annotator listed below; B – Percentage Agreement; C – Krippendorff's Alpha (α); D – Number of Agreements; E – Number of Disagreements; F – Number of Cases; G – Number of Decisions

Table 3: Agreement between Annotators

As seen in Table 3, the percentage agreement with annotators varied from 85% to 95% with

Krippendorff's Alpha (α) value achieving the least value of 0.76 and a maximum value of 0.907, respectively. As seen in Table 3, the number of sentences annotated by annotators varied from a minimum of 44 to a maximum of 347. This is due to the number of articles annotated by individual annotators.

The annotators in our study were requested to annotate any number of articles depending on their availability. While some chose to annotate a single article, three of the annotators (Annotators 1, 2 and 3 – shown in grey in Table 3) annotated three articles. The α value for these annotators was of the order 0.85 to 0.90. This shows that the increase in annotated sentences resulted in better agreement indicating the ease of applying the labels to sentences by these annotators.

The agreement achieved for each article between three of the annotators is tabulated in Table 4.

7 Conclusion

We presented in this paper an annotation scheme of context types for scientific articles, considering the persuasive characteristic of citations. We described the application of the Toulmin model for developing an argumentation framework for scientific articles, which was used for defining our context types. We discussed the results of the inter-rater reliability study carried out for establishing the reliability of our scheme. As we mentioned in Section 1, studies have successfully used this annotation scheme for developing tools that provide intelligent citation context based information services, indicating the usefulness of the annotation scheme.

Our future work involves examining the application of annotation schemes across other disciplines. We also intend to focus on using our context types for analyzing sentiments associated with citation contexts.

Article	A	B	C	D	E
Article 3	82.83	82.17	90.09	76.23	0.75
Article 2	84.73	86.74	90.36	77.10	0.77
Article 6	89.09	85.45	97.27	54.54	0.78
Article 7	90.47	95.23	90.47	58.71	0.82
Article 8	87.87	88.63	93.18	81.81	0.83
Article 4	90.21	85.32	91.74	93.57	0.84
Article 9	89.43	97.80	95.12	85.36	0.85
Article 5	93.93	91.81	85.45	94.54	0.87
Article 1	95.02	96.31	96.31	92.63	0.91

A – Average Pairwise percent agreement; B – Agreement between Annotator A and Annotator 2; C – Agreement between Annotator A and Annotator 1; D – Agreement between Annotator 1 and Annotator 2; E - Krippendorff's Alpha (α)

Table 4: Agreement for Articles

Classification Results				Confusion Matrix											
Label*	P	R	F		BGR	CWIG	CWO	IRCW	CWOG	DES	GAPS	ISSUE	FW	RWCW	UOCW	TOTAL
BGR	0.50	1.00	0.66	BGR	3	0	0	0	0	0	0	0	0	0	0	**3**
CWIG	1.00	1.00	1.00	CWIG	0	4	0	0	0	0	0	0	0	0	0	**4**
CWO	0.87	1.00	0.93	CWO	0	0	14	0	0	0	0	0	0	0	0	**14**
IRCW	0.92	1.00	0.96	IRCW	0	0	0	52	0	0	0	0	0	0	0	**52**
CWOG	1.00	0.66	0.80	CWOG	0	0	0	4	8	0	0	0	0	0	0	**12**
DES	0.85	0.87	0.86	DES	2	0	1	0	0	126	1	14	0	0	0	**144**
GAPS	0.95	0.87	0.91	GAPS	0	0	0	0	0	3	21	0	0	0	0	**24**
ISSUE	0.39	0.31	0.34	ISSUE	1	0	1	0	0	18	0	9	0	0	0	**29**
FW	1.00	1.00	1.00	FW	0	0	0	0	0	0	0	0	3	0	0	**3**
RWCW	1.00	1.00	1.00	RWCW	0	0	0	0	0	0	0	0	0	10	0	**10**
UOCW	1.00	1.00	1.00	UOCW	0	0	0	0	0	0	0	0	0	0	5	**5**
					6	**4**	**16**	**56**	**8**	**147**	**22**	**23**	**3**	**10**	**5**	**300**

* Labels CCW and SCCW are not shown in the table as none of the sentences were labeled with these labels.
Captions: P – Precision; R – Recall; F – F-Score

Table 1: Results of Pilot Study for each Label

References

Angrosh, M A, Cranefield, S., and Stanger, N. (2012a). Contextual Information Retrieval in Research Articles: Semantic Publishing Tools for the Research Community. Semantic Web Journal. Accepted for publication.

Angrosh, M.A., Cranefield, S., and Stanger, N. (2012b). Context Identification of Sentences in Research Articles: Towards Developing Intelligent Tools for the Research Community. Natural Language Engineering. DOI: 10.017/S1351324912000277.

Artstein, R., and Poesio, M. (2008). Survey Article Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4), 555-596.

Baldi, S. (1998). Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model. American Sociological Review, 63(6), 829. doi:10.2307/2657504

Bird, S., Loper, E. and Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. Journal of the American Society for Information Science, 36(4), 223-229. doi:10.1002/asi.4630360402

Brooks, T. A. (1986). Evidence of complex citer motivations. Journal of the American Society for Information Science, 37(1), 34-36. DOI:10.1002/asi.4630370106

Chubin, D. E., and Moitra, S. D. (1975). Content Analysis of References: Adjunct or Alternative to Citation Counting? Social Studies of Science, 5(4), 423-441.

Cozzens, S. E. (1989). What do citations count? The rhetoric-first model. Scientometrics, 15(5-6), 437-447. DOI:10.1007/BF02017064

Frost, C. O. (1979). The Use of Citations in Literary Research: A Preliminary Classification of Citation Functions. The Library Quarterly, 49(4), 399-414.

Gao, L., Tang, Z., and Lin, X. (2009). CEBBIP: A Parser of Bibliographic Information in Chinese Electronic Books. ACM/IEEE Joint Conference on Digital Libraries, JCDL 2009 (pp. 73-76). ACM Press.

Garfield, E. (1964). Science Citation Index - A new dimension in indexing. Science, 144(3619), 649-654.

Gilbert, G. N. (1977). Referencing as Persuasion. Social Studies of Science, 7(1), 113-122.

Hyland, K. (2002). Directives: Argument and Engagement in Academic Writing. Applied Linguistics, 23(2), 215-239. DOI:10.1093/applin/23.2.215

Ibekwe-sanjuan, F., Chen, C., and Pinho, R. (2007). Identifying Strategic Information from Scientific Articles through Sentence Classification. Journal of Applied Linguistics, 1518-1522.

Krippendorff, K. (2011). Computing Krippendorff's Alpha Reliability. Annenberg School of Communication. Departmental Papers (ASC). University of Pennsylvania. http://www.asc.upenn.edu/usr/krippendorff/mwebreliability4.pdf

Langer, H., Lüngen, H., and Bayerl, P. S. (2004). Text type structure and logical document structure. Proceedings of the 2004 ACL Workshop on Discourse Annotation - DiscAnnotation'04 (pp. 49-56). Morristown, NJ, USA: Association for Computational Linguistics. DOI:10.3115/1608938.1608945

Le, M.H., Ho, T.B., and Nakamori, Y. (2006). Detecting Citation Types Using Finite-State. PAKDD 2006, Lecture Notes in Artificial Intelligence 3918 (pp. 265-274). Springer-Verlag, Berlin Heidelberg.

Liakata, M. (2010). Zones of Conceptualisation in Scientific Papers: a Window to Negative and Speculative Statements. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (pp. 1-4).

Lipetz, B.-A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. American Documentation, 16(2), 81-90. DOI:10.1002/asi.5090160207

Majeed, A., Razak, S., Abu-Ghazaleh, N. B., & Harras, Kh. A. (2009). TCP over Multi-Hop Wireless Networks: The Impact of MAC Level Interactions. ADHOC-NOW 2009, Lecture Notes in Computer Science 5793 (pp. 1-15). Springer-Verlag, Berlin Heidelberg.

Mizuta, Y., and Collier, N. (2004). An Annotation Scheme for a Rhetorical Analysis of Biology Articles. Proceedings of the Fourth International Conference on Language Resource and Evaluation (LREC 2004) (pp. 1737-1740).

Moravcsik, M. J., and Murugesan, P. (1975). Some Results on the Function and Quality of Citations. Social Studies of Science, 5(1), 86-92. DOI:10.1177/030631277500500106

Nanba, H., and Okumura, M. (1999). Towards Multi-paper Summarization Retrieval of Papers Using Reference Information. In T. Dean (Ed.), IJCAI (pp. 926-931). Morgan Kaufmann.

Pham, S. B., and Hoffmann, A. (2003). A New Approach for Scientific Citation. In T. D. Gedeon and L. C. C. Fung (Eds.), Artificial Intelligence 2003 (pp. 759-771). Springer-Verlag Berlin Heidelberg.

Radoulov, R. (2008). Exploring Automatic Citation Classification. MSc. Thesis. University of Waterloo, Ontario.

Spiegel-Rosing, I. (1977). Science Studies: Bibliometric and Content Analysis. So-cial Studies of Science, 7(1), 97-113. DOI:10.1177/030631277700700111

Teufel, S. (1999). Argumentative Zoning: Information Extraction from Scientific Text. University of Edinburgh.

Teufel, S., and Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. Computational Linguistics. 28(4), 409-445.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006). Automatic classification of citation function. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006) (pp. 103-110). Association for Computational Linguistics.

Toulmin, S. (1958). The Uses of Argument. Cambridge University Press, Cambridge, England.

White, H. D. (2004). Citation Analysis and Discourse Analysis Revisited. Applied Linguistics, 25(1), 89-116. DOI:10.1093/applin/25.1.89

Wilbur, W. J., Rzhetsky, A., and Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC bioinformatics, 7, 356. DOI:10.1186/1471-2105-7-356