Contextual Information Extraction in Research Articles: A Case of developing contextual RDF data for ESWC papers

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

ABSTRACT
This paper reports our research work carried out for developing intelligent information systems using citation context information extracted from research articles. We explain in this paper the steps followed for identifying, extracting and managing contextual data from articles published in ESWC series. Reporting on the amount of triplification data produced in the process, we describe our application developed for providing value added information services for the research community.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing, Information Search and Retrieval, Digital Libraries

General Terms
Algorithms, Experimentation, Ontologies, Semantic Web

Keywords
Sentence Classification, Citation Classification, Sentence Context Ontology, Semantic Web

1. BACKGROUND
Research articles have emerged as an important medium of research communication. However, in recent times, a drastic increase in the research output has drawn significant research attention focused on developing intelligent information systems using the content. One of the prominent research questions has been to look into the task of extracting contextual information from the research content and employ these contexts for providing value added information services. However, the unstructured format of research content poses serious challenges in achieving this objective. Against this backdrop, the present research is taken up for developing intelligent information systems, exploiting the research content. After achieving good results with our initial work on ontologically modelling contexts of sentences in related work sections of research articles [1], we explored the possibility of identifying contexts of sentences across the entire article. This paper explains our research work carried out for extracting contextual data from articles published in the European Semantic Web Conference (ESWC) series. We report on the triplification data resulting from our research work and also describe the linked data application developed for using the derived contextual data.

I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria

2. WHAT WERE THE CONTEXT TYPES DEFINED IN THE STUDY?
We considered only those paragraphs containing citations and classified sentences in each paragraph into citation sentences and non-citation sentences. We defined a classification scheme of ten context types for citation sentences and seven context types for non-citation sentences. The context types are shown in Figure 1. These were defined after manually analyzing 331 citation sentences and 838 non-citation sentences respectively from our training dataset of 20 articles selected from the Lecture Notes in Computer Science (LNCS) collection at springerlink.com. We proposed a framework based on a generic rhetorical pattern observed in paragraphs of the training dataset as shown in Figure 1. The contexts served as our labels for sentences that were used in machine learning experiments and also provided a basis for developing the Sentence Context Ontology. More details about the context types and the framework are presented by Angrosh et al. [2].

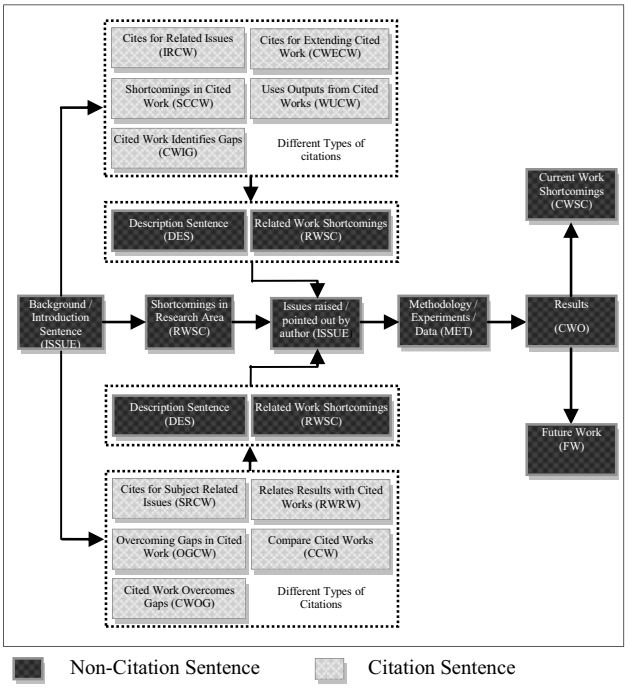

Figure 1: Context types for sentences in research articles contained in paragraphs with citation sentences

3. HOW DID WE MODEL THE CONTEXTUAL INFORMATION?
In order to model the different contexts identified above with suitable relations, we developed the Sentence Context Ontology which facilitated in deriving RDF data. The ontology is available

at https://info-nts-12.otago.ac.nz:8090/sentcon/. The ontology imports the Bibliographic Ontology with namespace http://purl.org/. A schematic diagram of the ontology is shown in Figure 2.

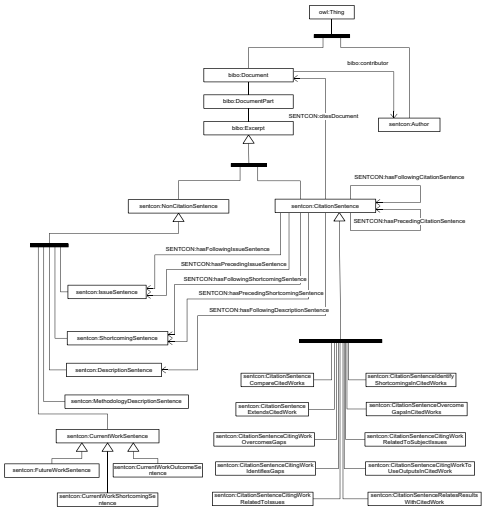

Figure 2: Sentence Context Ontology

4. HOW DID WE ACHIEVE AUTOMATIC CONTEXT IDENTIFICATION?

We carried out supervised learning experiments using Maximum Entropy Markov Models (MEMMs) [3] and Conditional Random Fields (CRFs) [4] for achieving automatic context identification. We defined various features and tested their ability to predict sentence context labels using our training dataset. The results of our experiments are tabulated in Table 1. More details of the machine learning experiments are presented by Angrosh et al. [2].

Table 1: Results of the Classifier

Label	Accuracy: 93.37% CRF – 1st & 0 Order			Accuracy: 89.58% CRF – 1st Order			Accuracy: 68.33% MEMM		
	P	R	F	P	R	F	P	R	F
FW	1.00	0.96	0.98	1.00	0.93	0.96	1.00	0.64	0.78
CWO	0.96	1.00	0.98	0.92	0.90	0.91	0.00	0.00	0.00
ISSUE	0.97	0.97	0.97	0.93	0.96	0.95	0.79	0.93	0.86
METH	0.91	0.98	0.94	0.90	0.92	0.91	0.57	0.66	0.61
WUCW	0.94	0.92	0.93	0.90	0.94	0.92	0.86	0.84	0.85
SRCW	0.88	0.98	0.92	0.84	0.91	0.88	0.78	0.65	0.70
IRCW	0.89	0.96	0.92	0.81	0.95	0.87	0.31	0.54	0.39
RWSC	0.90	0.94	0.92	0.86	0.94	0.90	0.75	0.85	0.80
DES	0.89	0.89	0.89	0.85	0.86	0.85	0.75	0.59	0.66
CWIG	0.91	0.68	0.78	1.00	0.50	0.66	0.66	0.12	0.21
SCCW	0.86	0.67	0.76	0.77	0.50	0.60	1.00	0.14	0.25
CWOG	0.77	0.73	0.75	0.66	0.42	0.51	1.00	0.10	0.19
CWECW	1.00	0.50	0.66	1.00	0.25	0.40	0.00	0.00	0.00
CWSC	0.62	0.45	0.52	0.66	0.18	0.28	0.00	0.00	0.00
RWRW	1.00	0.12	0.22	0.00	0.00	0.00	0.00	0.00	0.00
CCW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

P – Precision, R – Recall, F- F-Score

5. HOW DID WE EXTRACT INFORMATION FROM ARTICLES?

Various modules were developed in Python for identifying, extracting and managing information from research articles. The architecture of the system is shown in Figure 3.

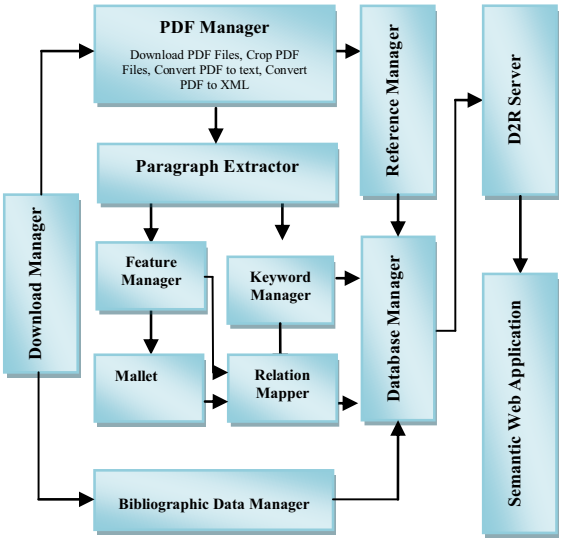

Figure 3: Architecture of our Information Extraction System

6. HOW MUCH DATA HAVE WE GENERATED?

Employing the techniques described above, we extracted information from research articles published in the European Semantic Web Conference (ESWC). Table 2 provides the details of the extracted data from the ESWC series. As seen from the table, presently, we have extracted data from seven volumes published in the ESWC series. This provided a total of 355 articles♠. Paragraphs with citation sentences were extracted from these articles and a total of 10,564 sentences were extracted from these paragraphs. This included a total of 6,158 citation sentences and 4,406 non-citation sentences respectively. Further, a total of 17,844 authors were extracted from cited works (citations) in these articles and a total number of 4,121 documents were cited by these articles.

Table 2: Details of Sentences extracted from ESWC volumes

Year	Volume	Articles	Sentences		Citations	
			CS	NCS	A	CD
2005	3552	42	805	570	2455	874
2006	4011	45	896	660	2277	844
2007	4519	48	1087	544	2612	902
2008	5021	66	871	795	3198	1102
2009	5554	76	1214	867	3513	1161
2010	6088	27	538	406	1564	519
2010	6089	51	747	564	2225	723
Total		**355**	**6158**	**4406**	**17844**	**6125**
Total Number of Sentences: 10564						

CS – Citation Sentences; NCS – Non-Citation Sentences; A – Authors; CD – Cited Documents

The details of different types of citation and non-citation sentences extracted from each volume are provided in Table 3 and 4 respectively.

♠ A few articles published in these seven volumes were not considered for different reasons. Some were invited talks, while some followed a different reference format.

Table 3: Details of Citation Sentences extracted from ESWC

Year	LNCS Vol.	Citation Sentences									
		A	B	C	D	E	F	G	H	I	J
2005	3552	286	355	31	8	87	0	32	0	6	-
2006	4011	299	375	10	56	82	0	63	0	10	1
2007	4519	402	416	24	49	112	0	64	0	17	3
2008	5021	306	336	28	39	83	0	62	0	11	6
2009	5554	463	438	30	66	137	0	61	0	15	4
2010	6088	231	167	7	37	48	0	38	0	7	3
2010	6089	347	240	10	41	55	0	46	0	7	1
Total		**2334**	**2327**	**140**	**296**	**604**	**0**	**366**	**0**	**73**	**18**
Total Number of Citation Sentences: 6158											

A – Citation Sentence Cites Works Related to Issues; B – Citation Sentence Cites Works Related to Subject Issues; C – Citation Sentence Cites Works Identifying Gaps; D – Citation Sentence Cites Works Overcoming Gaps; E – Citation Sentence Identifies Shortcomings in Cited Work; F – Citation Sentence Extends Current Cited Work; G – Citation Sentence Uses Outputs in Cited Work; H– Citation Sentence Overcome Gaps in Cited Work; I – Citation Sentence Compares Results to Cited Work; J – Citation Sentence Compares Cited Works

Table 4: Details of Non-Citation Sentences extracted from ESWC volumes

Year	LNCS Volume	Non-Citation Sentences					
		A	B	C	D	E	F
2005	3552	159	237	101	32	11	30
2006	4011	162	282	128	45	12	31
2007	4519	126	240	104	40	9	25
2008	5021	206	360	138	43	9	39
2009	5554	206	386	157	52	17	49
2010	6088	91	177	76	27	8	27
2010	6089	139	247	98	40	10	30
	Total	**1089**	**1929**	**802**	**279**	**76**	**231**
Total Number of Non-Citation Sentences: 4406							

A – Description sentences
B – Shortcoming sentences
C – Current Work Outcome
D – Current Work Shortcoming Sentence
E – Future Work Sentence
F – Methodology Description Sentence

The extracted data were stored in relation form and were converted to RDF data using the D2R Server. The mapping file of the D2R Server was configured as per the Sentence Context Ontology. This resulted in about 250,000 triples.

7. WHAT APPLICATION IS DEVELOPED USING TRIPLIFICATION DATA?

Currently, we have developed a linked data application using the triplification data for providing value added information services for the research community. The application is available at https://info-nts-12.otago.ac.nz:8090/cirrademo/. It uses SPARQL to query RDF data and Exhibit Timeline for providing interactive user interfaces. The following sections briefly explain the features and different timelines provided by the application.

7.1 Use of Data from a SPARQL Endpoint

The application uses DBLP linked data available at the SPARQL endpoint – http://rkbexplorer.com for retrieving metadata of articles in ESWC using keywords extracted from these titles.

7.2 Citation Sentences Timeline

The application supports viewing all citation sentences along with their contexts on a timeline. The timeline is an interactive interface that enables horizontal scrolling of information placed on the timeline. The citation sentences of an article are placed according to the year of the cited work and are distinguished by the use of different colours, with each colour signifying a different context type. Users can select citation sentences of a specific type. For example, if an user is interested in viewing only those citation sentences that identify shortcomings in the cited work, he/she can accordingly choose that context type to view only those citation sentences. Clicking on one of these citation sentences, results in the display of associated sentences in a lens view. These include (a) preceding issue sentence; (b) preceding shortcoming sentence; (c) following description sentence; (d) following issue sentence; (e) following shortcoming sentence; (f) preceding citation sentence and (g) following citation sentence. The Sentence Context Ontology is used to model this information. Furthermore, each citation sentence displays cited work, which the user can use to navigate to the citations timeline to learn more about the cited work. The timeline also facilitates navigation to the author timeline for learning more about an author's work.

7.3 Author Timeline

The author timeline displays all works of an author by distinguishing between citing and cited works along with the contexts.

7.4 Citations Timeline

The citations timeline displays the contexts in which a specific cited work is used by various authors in different articles on the timeline. This forms an important tool for the research community as it helps in better understanding of the cited work by easily viewing different contexts in which the cited work was used.

7.5 Keywords Timeline

The keywords timeline allows for searching citation sentences using keywords. The application extracts keywords from citation sentences and creates an index for this purpose. This facilitates in sketching the intellectual lineage for ideas by understanding how different people have cited different works for specific keywords.

7.6 Export Data in Different Formats

The application allows for exporting data in different formats such as RDF/XML, Semantic wikitext, Exhibit JSON and tab separated values.

More details of the application are described by Angrosh et al. [2].

8. REFERENCES

[1] M.A. Angrosh, S. Cranefield, and N. Stanger, "Ontology-based Modelling of Related Work Sections in Research Articles: Using CRFs for Developing Semantic Data based Information Retrieval Systems," Proceedings of the 6th International Conference on Semantic Systems (I-Semantics), ACM, 2010.

[2] M.A. Angrosh, S. Cranefield, and N. Stanger, "Contextual Information Retrieval in Research Articles: Semantic Publishing Tools for the Research Community," Discussion paper 2011/06, Department of Information Science, University of Otago, Dunedin, New Zealand.

[3] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," Proceedings of the International Conference on Machine Learning, 2000, pp. 591-598.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data," Proceedings of the International Conference on Machine Learning, 2001, pp. 282-289.