Applying Core Scientific Concepts to Context-Based Citation Recommendation

Daniel Duma[1], Maria Liakata[2], Amanda Clare[3], James Ravenscroft[2], Ewan Klein[1]
University of Edinburgh[1], University of Warwick[2], University of Aberystwyth[3]
danielduma@gmail.com, m.liakata@warwick.ed.ac.uk, afc@aber.ac.uk, ravenscroft@papro.org.uk, ewan@inf.ed.ac.uk

Abstract

The task of recommending relevant scientific literature for a draft academic paper has recently received significant interest. In our effort to ease the discovery of scientific literature and augment scientific writing, we aim to improve the relevance of results based on a shallow semantic analysis of the source document and the potential documents to recommend. We investigate the utility of automatic argumentative and rhetorical annotation of documents for this purpose. Specifically, we integrate automatic Core Scientific Concepts (CoreSC) classification into a prototype context-based citation recommendation system and investigate its usefulness to the task. We frame citation recommendation as an information retrieval task and we use the categories of the annotation schemes to apply different weights to the similarity formula. Our results show interesting and consistent correlations between the type of citation and the type of sentence containing the relevant information.

Keywords: context-based, citation recommendation, CoreSC, classification, annotation

1. Introduction

Given the need to navigate the ever-increasing volume of scientific literature, the task of *Context-Based Citation Recommendation* (CBCR) has recently received a lot of interest. The task consists in recommending relevant papers to be cited at a specific point in a draft scientific paper, and is universally framed as an information retrieval scenario. However, in order to make these suggestions as useful and relevant as possible, we argue here that we need to apply a measure of understanding to the text of the draft paper.

To this end, we investigate the applicability of rhetorical annotation schemes for this task. A number of different schemes for scientific papers have been proposed over the years and several of them have yielded annotated resources, which enable training algorithms for automatic annotation. Like others before, we evaluate our performance at this task by trying to recover the original citations found in papers that have already been published.

In an information retrieval scenario like ours, a list of documents ranked by relevance is returned in reply to a query, aiming to satisfy the user's information need. Our hypothesis is that if we can a) classify the information need into a number of discrete categories, and b) classify each sentence in a document according to its function or contribution to it, we could use this segmentation of text to increase the relevance of recommendations.

2. CBCR: Previous Work

The CBCR task considers that we need to recommend a citation for each *citation placeholder*: a special token inserted in the text of a draft paper where the citation should appear. In a standard Information Retrieval (IR) approach, the corpus of potential papers to recommend (the *document collection*) is indexed for retrieval using a standard vector-space-model approach. Then, for each citation placeholder, the textual context around it (the *citation context*) is treated as the *query*, and a similarity measure is applied to rank the documents in the collection.

Perhaps the seminal piece of work in this area is He et al. (2010), who built an experimental citation recommendation system using the documents indexed by the CiteSeerX search engine as a test collection (over 450,000 documents), now available for testing online (Huang et al., 2014). Recently, Huang et al. (2015) improved all metrics on this task and dataset by applying multi-layered neural networks. Other techniques have been applied to this task, such as collaborative filtering (Caragea et al., 2013) and translation models (He et al., 2012), and other aspects of it have been explored, such as document representation (Duma and Klein, 2014) and context extraction (Ritchie, 2009).

The work we present here is to our knowledge the first to apply the classification of rhetorical function of sentences to the CBCR task.

3. Classification of Rhetorical Function

Scientific papers follow a formal structure, and the language of academia requires clear argumentation (Hyland, 2009). This has led to the creation of classification schemes for the rhetorical and argumentative structure of scientific papers. To date, the standard approach has been to take the sentence as the minimum unit of annotation, and we maintain this approach in this work.

Two of the most prominent are Argumenative Zoning (Teufel, 2000) and Core Scientific Concepts (CoreSC, Liakata et al. (2010)). These are among the first approaches to incorporate successful automatic classification of sentences in a paper, using a supervised machine learning approach.

CoreSC (Table 1) was specifically developed for the domain of biomedical science and treats papers as "human-readable representations of scientific investigations", aiming to retrieve the structure of the investigation from the paper (Liakata et al., 2010).

Rhetorical and argumentation schemes like these have found application in experimental academic retrieval tools (Schäfer and Kasterka (2010), Ravenscroft et al. (2013), Angrosh et al. (2013)) and here we explore their potential application to a deeper integration with the writing process.

For the task of recommending a citation for a given span of text, the ideal resource for classifying these spans would

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-Old	A method mentioned pertaining to previous work
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	The data/phenomena recorded in an investigation
Result	Factual statements about the outputs of an investigation
Conclusion	Statements inferred from observations & results relating to research hypothesis

Table 1: CoreSC classes and their description.

deal with the *function* of a citation within its argumentative context. Specific schemes for classifying the function of a citation have been developed, notably that of Teufel et al. (2006), specifically developed for Citation Function Classification. However, we are not aware of a scheme particularly tailored to our domain of biomedical science, so instead we employ CoresSC classes as proxies for citation function, which we hypothesize is valid in our domain.

4. Methodology

Given that we can classify each sentence according to its rhetorical status in the document using Core Scientific Concepts, we aim to find whether a) giving higher weight to terms found in particular classes of sentences in the document collection will increase the retrieval accuracy and b) whether we can increase it further by correlating this weighting with the function of the citation that we are trying to recommend papers for.

This is our methodology:

1. Firstly, we automatically label the sentences in each document in our collection using CoreSC.
2. We then index these documents and for each document we create a separate field for each class of sentence. We index all sentences of the same class into the same field for each document. That is, we index a bag-of-words for each class (Hypothesis, Background, Method, etc.), which contains all the words from all sentences of that class present in the document.
3. We label the *insertion context* with a rhetorically-motivated class that encodes the citation function. In our implementation, this is just the class of the sentence as classified in the previous step, so a citation's type is the CoreSC class of the sentence containing it.
4. We test different weights for each citation function (as labelled in step 3) and compare the results with the baseline of using all weights equally set to 1. The evaluation method is described below. We evaluate by performing K-fold cross-validation and comparing the results over all fold combinations. What we expect to find is not only improvement on average in the scores for a particular citation type, but consistency across folds in weights and obtained improvement.

Our hypothesis is that the relevance of suggested citations can be significantly increased by applying a set of automatically-trained per-field weights to the similarity function.

We try to find the best combination of weights to set for each citation function. For example, we may find that for all citations of type Method-Old, if we gave higher weight to the content of sentences of type Method-New in the documents in the collection, we would achieve a higher accuracy (see Figure 1 for an illustration).

Evaluation: We aim to reduce purpose-specific annotation, so we evaluate the performance of our recommendation against existing scientific publications. We substitute all citations in the text with *citation placeholders* and make it our task to match each placeholder with the correct reference that was originally cited. We only consider *resolvable citations*, that is, citations to references that point to a paper that is in our collection, which means we have access to its full machine-readable contents (*collection-internal references*).

The task then becomes, for each citation placeholder: 1) to extract its context, and from it a query, and 2) attempt to retrieve the original paper cited in the context from the whole document collection. We measure how well we did at our task by how far down the list of ranked retrieval results we find the original paper cited. We use two metrics to measure accuracy: Normalized Discounted Cumulative Gain (NDCG), a smooth discounting scheme over ranks, and top-1 accuracy, which is just the number of times the original paper was retrieved in the first position.

Context extraction: Lacking a more sophisticated method, we extract the context of a citation using a symmetric window of 3 sentences: 1 before the citation, the sentence containing the citation and 1 after. This is a frequently applied method (Huang et al., 2015) and is close to what has been assumed to be the optimal window of 2 sentences up, 2 down (Qazvinian and Radev, 2010), while yielding fewer query terms and therefore allowing us more experimental freedom through faster queries.

Similarity: We use the default Lucene similarity formula for assessing the relevance of a document to a context (Figure 2).

In this formula, the coord term is an absolute multiplier of

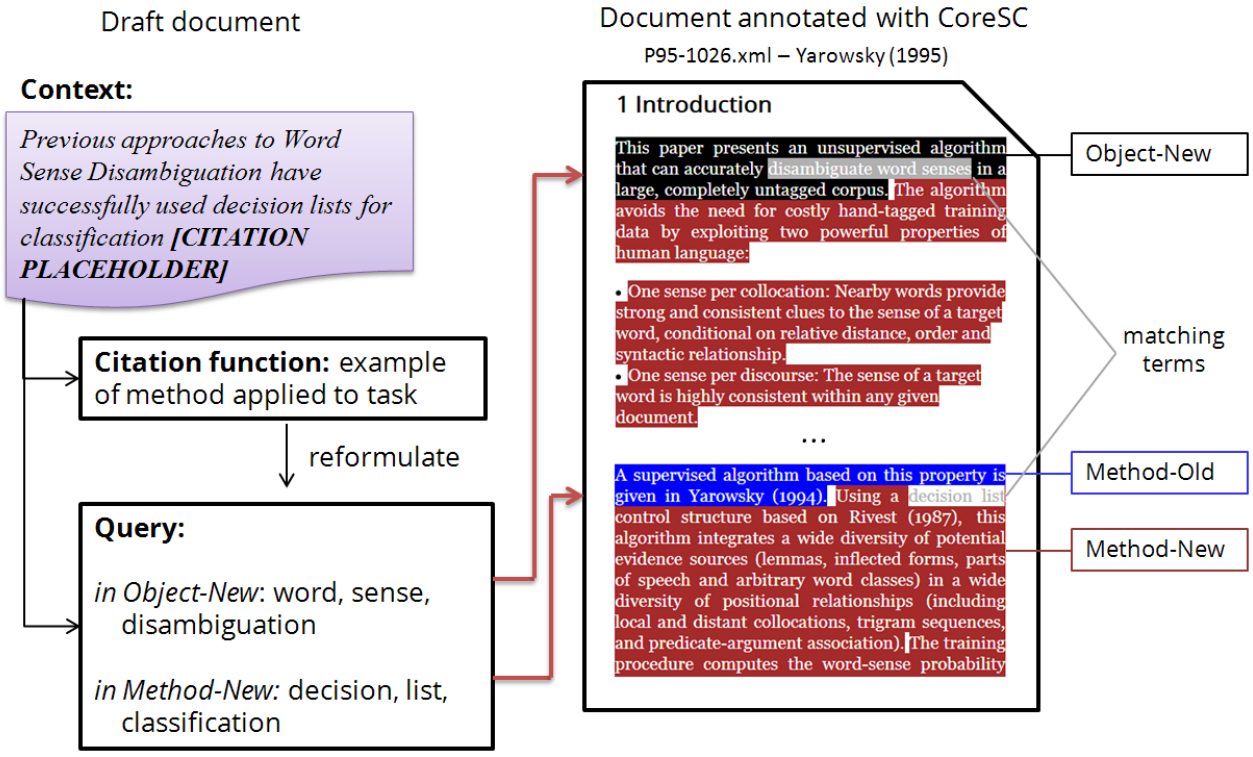

Figure 1: The intuition behind our approach. Depending on the function of the citation, we search for key terms in different classes of sentences, as automatically labelled. In practice, it is more finely grained by applying different weights to different document fields.

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \sum_{t \in q} \text{tf}(t \in d) \cdot \text{idf}(t)^2 \cdot \text{norm}(t,d)$$

Figure 2: Default Lucene similarity formula

the number of terms in the query q found in the document d, tf is the absolute term frequency score of term t in document d, $\text{idf}(t)$ is the inverse document score and norm is a normalization factor that divides the overall score by the length of document d. Note that all these quantities are per-field, not per-document.

Technical implementation: We index the document collection using the Apache Lucene retrieval engine, specifically through the helpful interface provided by elasticsearch 2.1[1]. For each document, we create one field for each CoreSC class, and index into each field all the words from all sentences in the document that have been labelled with that class.

The *query* is formed of all the terms in the citation's context that are not in a short list of stopwords. Lucene queries take the basic form *field:term*, where each combination of *field* and *term* form a unique term in the query. We want to match the set of extracted terms to all fields in the document, as each field represents one class of CoreSC.

The default Lucene similarity formula (Figure 2) gives a boost to a term matching across multiple fields, which in our case would introduce spurious results. To avoid this, we employ DisjunctionMax queries, where only the top scoring result is evaluated out of a number of them. Having one query term for each of the classes of CoreSC for each distinct token (e.g. *Bac:"method"*, *Goa:"method"*, *Hyp:"method"*, etc.), only the one with the highest score will be evaluated as a match.

Weight training: Testing all possible weight combinations is infeasible due to the combinatorial explosion, so we adopt the greedy heuristic of trying to maximise the objective function at each step.

Our weight training algorithm can be summarized as "hill climbing with restarts". For each fold, and for each citation type, we aim to find the best combination of weights to set on sentence classes that will maximise our metric, in this case the NDCG score that we compute by trying to recover the original citation. We keep the queries the same in structure and term content and we only change the weights applied to each field in a document to recommend. Each field, as explained above, contains only the terms from the sentences in the document of one CoreSC class.

The weights are initialized at 1 and they move by -1, 6, and -2 in sequence, going through a minimum of 3 iterations. Each time a weight movement is applied, it is only kept if the score increases. Otherwise the previous weight value is restored.

This simple algorithm is not guaranteed to find a globally optimal combination of parameters for the very complex function we are optimizing, but it is sufficient for our current objective. We aim to explore other machine learning techniques for learning weights in future work.

[1] https://elastic.co/

Type	Citations	Fold 1	Fold 2	Fold 3	Fold 4	Folds improved	% impr.	Std. Dev.
Con	133	6.63	16.96	4.24	9.84	4	9.42	5.53
Bac	700	28.77	7.56	28.32	24.44	4	22.27	10.00
Met	602	23.70	43.78	18.44	19.28	4	26.30	11.88
Res	178	33.87	54.18	7.28	30.95	4	31.57	19.21
Goa	44	73.74	46.47	32.38	28.40	4	45.25	20.52
Obj	65	547.57	19.32	91.31	18.30	4	169.13	254.60
Mod	161	-3.64	17.59	32.66	18.33	3	16.24	14.96
Obs	17	-18.78	37.06	9.57	102.40	3	32.56	51.84
Exp	16	-23.04	46.50	105.23	31.63	3	40.08	52.73
Hyp	31	43.14	-37.93	-38.43	6.03	2	-6.80	39.28
Mot	8	0	0	0	0	0	0.00	0.00

Figure 3: Results of evaluating with 4-fold cross-validation by citation type, ordered by number of folds showing improvement and by standard deviation. The citation type is the CoreSC class of the sentence containing the citation.In bold, citation types for which there was improvement across all folds.

			Weights											Scores		
Type	Citations	Fold	Bac	Con	Exp	Goa	Hyp	Met	Mod	Mot	Obj	Obs	Res	NDCG*	Accuracy*	% imp.
Bac	700	1	1	1	1	0	0	1	1	0	0	0	1	0.401	0.183	28.77
		2	1	0	1	0	0	1	0	0	0	1	1	0.440	0.160	7.56
		3	1	1	0	0	0	1	1	0	0	1	1	0.344	0.109	28.32
		4	1	1	1	0	0	1	1	0	0	0	1	0.437	0.143	24.44
Con	133	1	1	1	1	0	0	0	0	0	0	0	1	0.214	0.059	6.63
		2	1	1	0	0	0	1	1	0	0	0	0	0.528	0.212	16.96
		3	1	0	1	0	0	1	0	0	0	0	0	0.490	0.242	4.24
		4	1	1	1	0	0	1	1	0	0	0	0	0.339	0.152	9.84
Goa	44	1	1	0	1	0	0	1	0	0	0	0	0	0.459	0.182	73.74
		2	1	0	0	0	0	1	0	0	0	0	0	0.126	0.000	46.47
		3	1	0	0	0	0	1	0	0	0	1	0	0.309	0.182	32.38
		4	1	0	0	0	0	1	0	0	0	0	0	0.214	0.091	28.40
Met	602	1	1	0	0	0	0	1	0	0	0	0	1	0.404	0.146	23.70
		2	2	0	0	0	0	0	1	0	0	0	0	0.332	0.139	43.78
		3	2	0	0	0	0	1	0	0	0	0	1	0.467	0.173	18.44
		4	2	0	0	0	0	0	1	0	0	0	0	0.288	0.107	19.28
Obj	65	1	7	0	1	0	0	1	0	0	0	1	1	0.085	0.059	547.57
		2	7	0	1	1	0	1	0	0	0	1	1	0.122	0.063	19.32
		3	7	0	1	0	0	1	0	0	0	1	7	0.114	0.000	91.31
		4	13	0	1	1	0	4	0	0	0	1	0	0.235	0.063	18.30
Res	178	1	1	0	0	0	0	0	0	0	0	1	1	0.379	0.111	33.87
		2	1	0	0	0	0	0	0	0	0	1	1	0.350	0.133	54.18
		3	1	0	0	0	0	0	0	0	0	0	1	0.638	0.273	7.28
		4	1	0	0	0	0	0	0	0	0	0	1	0.362	0.091	30.95

Figure 4: Weight values for the citation types that improved across all folds. The weight values for the 4 folds are shown, together with test scores and improvement over the baseline. The weight cells are shaded according to their value, darker is higher. In bold, citation types that consistently improve across folds. On the right-hand side are the scores obtained through testing and the percentage increase over the baseline, in which all weights were set to 1. *NDCG and Accuracy (top-1) are averaged scores over all citations in the test set for that fold.

5. Experiments

Our corpus is formed of 663000 papers from the PubMed Central Open Access corpus[2]. These papers are already provided in a clean, hand-authored XML format with a well-defined XML schema[3]. For our experiments we created our document collection out of the papers published up to and including 2013, and selected the top 100 documents with the most collection-internal references published in or after 2014 as our test set, from which we extract the citations and citation contexts.

We need to test whether our conditional weighting of text spans based on CoreSC classification is actually reflecting

[2]http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[3]http://jats.nlm.nih.gov/

some underlying truth and is not just a random effect of the dataset. To this end, we perform K-fold cross-validation on the corpus, where we learn the weights for K-1 folds and test their impact on one fold, and we report the averaged gains over each fold.
The full source code employed to run these experiments and instructions on how to replicate them are available on GitHub[4].

6. Results and Discussion

Figure 3 shows the results of evaluating with 4-fold cross-validation by citation type, ordered by number of folds improved and standard deviation. The citation type is the CoreSC class of the sentence containing the citation. We can see that 6 out of 11 types of citation exhibit improvement across all folds, and there is a relationship between the standard deviation of the improvement in scores and the number of citations of that type.
Figure 4 expands on this and shows the best weight values that were found for each fold, for all 6 citation types for which there was improvement on all folds. On the right-hand side are the scores obtained *after testing* and the percentage increase over the baseline, in which all weights are set to 1.
As is to be expected, the citations are skewed in numbers towards some CoreSC classes. A majority of citations occur within sentences that were automatically labelled Background and Methodology, no doubt due to a pattern in the layout of the content of articles. This yields many more Bac and Met citations to evaluate on, and for this reason we set a hard limit to the number of citations per zone to 700 in these experiments.
These are initial results and as such should be treated with caution. This said, a number of patterns are immediately evident. For all citation functions, it seems to be universally useful to know that the candidate document matches the query term in sentences from its Background sentences or Method sentences. It is also possible that this is partly an effect of there being more sentences of type Background and Method in a candidate paper.
Similarly, it seems it is better to ignore other classes of sentences in candidate papers, such as Motivation and Observation. Note here that the fact that a weight combination was found where the best weight for a CoreSC class is 0 does not mean that including information from this zone is not useful but rather that it is in fact *detrimental*, as eliminating it actually increased the NDCG score.
Interestingly, for citations in Results sentences, only Background, Results and, to a lesser degree Observations sentences in candidate documents seem to contain useful information. This is not surprising, and it allows for easy interpretation: when reporting results, these are often compared with previous results reported in other papers.
The degree of consistency varies across citation types. For Bac, Con, Goa, Met, Obj and Res, improvements are found at each fold and it seems that some consistency can be found in the trained weights. These are also types with a significant number of citations available. Exp, Hyp, Mod and Obs are the ones that are inconsistent in improvement: for some folds, the trained weights actually decrease the score, which suggests that no clear pattern is to be found. These are generally classes with fewer citations available, which could go some way towards explaining this. However, the exception here is Mod, which, in spite of a significant number of citations (161), still exhibits inconsistency, with the first fold decreasing in score.
It is important to note that our evaluation pipeline necessarily consists of many steps, and encounters issues with XML conversion, matching of citations with references, matching of references in papers to references in the collection, etc., where each step in the pipeline introduces a degree of error that we have not estimated here. Perhaps the single most significant one is that of the automatic sentence classifier.
The performance of the Sapienta classifier[5] we employ here has recently been independently tested on a different corpus from the originally annotated corpus used to train it. It yielded 51.9% accuracy over all eleven classes, improving on the 50.4% 9-fold cross-validation accuracy over its training corpus (Ravenscroft et al., 2016).
Further to this, we judge that the consistency of correlations we find confirms that what we can see in Figure 4 is not due to random noise, but rather hints at underlying patterns in the connections between scientific articles in the corpus. This also seems to confirm our assumption that the CoreSC class of the sentence that a citation appears in can be used as a proxy for the function of this citation.

7. Conclusion and Future Work

We have presented a novel application of CoreSC rhetorical function classification to context-based citation recommendation, an information retrieval application.
We have found strong indications of correlation between different classes of sentences in citing and cited articles. This suggests that there are gains to be reaped in a practical application of CoreSC to context-based citation recommendation. However, more experiments are required to confirm these initial results, and it still remains to be evaluated versus more standard approaches. One key piece of future work will be including the study of "anchor text", that is, citations to a document found in other documents, which is a key source of information for the CBCR task.

8. References

Angrosh, M., Cranefield, S., and Stanger, N. (2013). Context identification of sentences in research articles: Towards developing intelligent tools for the research community. *Natural Language Engineering*, 19(04):481–515.

Caragea, C., Silvescu, A., Mitra, P., and Giles, C. L. (2013). Can't see the forest for the trees?: a citation recommendation system. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 111–114. ACM.

Duma, D. and Klein, E. (2014). Citation resolution: A method for evaluating context-based citation recommendation systems. In *Proceedings of the 52nd Annual*

[4]https://github.com/danieldmm/minerva

[5]http://www.sapientaproject.com

Meeting of the Association for Computational Linguistics (Short Papers), page 358ï¿½363, Baltimore, Maryland, USA.

He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM.

He, J., Nie, J.-Y., Lu, Y., and Zhao, W. X. (2012). Position-aligned translation model for citation recommendation. In *String Processing and Information Retrieval*, pages 251–263. Springer.

Huang, W., Wu, Z., Mitra, P., and Giles, C. L. (2014). Refseer: A citation recommendation system. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*, pages 371–374. IEEE.

Huang, W., Wu, Z., Liang, C., Mitra, P., and Giles, C. L. (2015). A neural probabilistic model for context based citation recommendation.

Hyland, K. (2009). *Academic discourse: English in a global context*. Bloomsbury Publishing.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. R. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *LREC*.

Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 555–564. Association for Computational Linguistics.

Ravenscroft, J., Liakata, M., and Clare, A. (2013). Partridge: An effective system for the automatic cassification of the types of academic papers. In *Research and Development in Intelligent Systems XXX*, pages 351–358. Springer.

Ravenscroft, J., Oellrich, A., Saha, S., and Liakata, M. (2016). Multi-label annotation in scientific articles -the multi-label cancer risk assessment corpus. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*.

Ritchie, A. (2009). Citation context analysis for information retrieval. Technical report, University of Cambridge Computer Laboratory.

Schäfer, U. and Kasterka, U. (2010). Scientific authoring support: A tool to navigate in typed citation graphs. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing: Writing processes and authoring aids*, pages 7–14. Association for Computational Linguistics.

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

Teufel, S. (2000). *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.