Automatic classification of citation function

Simone Teufel **Advaith Siddharthan** **Dan Tidhar**
Natural Language and Information Processing Group
Computer Laboratory
Cambridge University, CB3 0FD, UK
{Simone.Teufel,Advaith.Siddharthan,Dan.Tidhar}@cl.cam.ac.uk

Abstract

The automatic recognition of the rhetorical function of citations in scientific text has many applications, from improvement of impact factor calculations to text summarisation and more informative citation indexers. Citation function is defined as the author's reason for citing a given paper (e.g. acknowledgement of the use of the cited method). We show that our annotation scheme for citation function is reliable, and present a supervised machine learning framework to automatically classify citation function, which uses several shallow and linguistically-inspired features. We find, amongst other things, a strong relationship between citation function and sentiment classification.

1 Introduction

Why do researchers cite a particular paper? This is a question that has interested researchers in discourse analysis, sociology of science, and information sciences (library sciences) for decades (Garfield, 1979; Small, 1982; White, 2004). Many annotation schemes for citation motivation have been created over the years, and the question has been studied in detail, even to the level of in-depth interviews with writers about each individual citation (Hodges, 1972).

Part of this sustained interest in citations can be explained by the fact that bibliometric metrics are commonly used to measure the impact of a researcher's work by how often they are cited (Borgman, 1990; Luukkonen, 1992). However, researchers from the field of discourse studies have long criticised purely quantitative citation analysis, pointing out that many citations are done out of "politeness, policy or piety" (Ziman, 1968), and that criticising citations or citations in passing should not "count" as much as central citations in a paper, or as those citations where a researcher's work is used as the starting point of somebody else's work (Bonzi, 1982). A plethora of manual annotation schemes for citation *motivation* have been invented over the years (Garfield, 1979; Hodges, 1972; Chubin and Moitra, 1975). Other schemes concentrate on citation *function* (Spiegel-Rüsing, 1977; O'Connor, 1982; Weinstock, 1971; Swales, 1990; Small, 1982)). One of the best-known of these studies (Moravcsik and Murugesan, 1975) divides citations in running text into four dimensions: conceptual or operational use (i.e., use of theory vs. use of technical method); evolutionary or juxtapositional (i.e., own work is based on the cited work vs. own work is an alternative to it); organic or perfunctory (i.e., work is crucially needed for understanding of citing article or just a general acknowledgement); and finally confirmative vs. negational (i.e., is the correctness of the findings disputed?). They found, for example, that 40% of the citations were perfunctory, which casts further doubt on the citation-counting approach.

Based on such annotation schemes and hand-analyzed data, different influences on citation behaviour can be determined. Nevertheless, researchers in the field of citation content analysis do not normally cross-validate their schemes with independent annotation studies with other human annotators, and usually only annotate a small number of citations (in the range of hundreds or thousands). Also, automated application of the annotation is not something that is generally considered in the field, though White (2004) sees the future of discourse-analytic citation analysis in automation.

Apart from raw material for bibliometric studies, citations can also be used for search purposes in document retrieval applications. In the library world, printed or electronic citation indexes such as ISI (Garfield, 1990) serve as an orthogonal

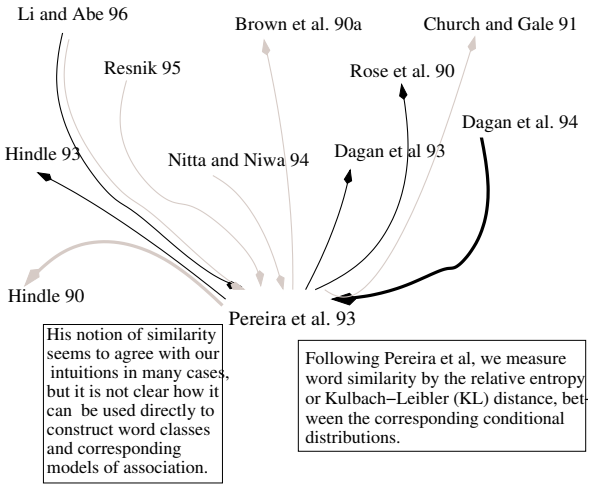

Figure 1: A rhetorical citation map

search tool to find relevant papers, starting from a source paper of interest. With the increased availability of documents in electronic form in recent years, citation-based search and automatic citation indexing have become highly popular, cf. the successful search tools Google Scholar and CiteSeer (Giles et al., 1998).[1]

But not all search needs are fulfilled by current citation indexers. Experienced researchers are often interested in *relations* between articles Shum (1998). They want to know if a certain article criticises another and what the criticism is, or if the current work is based on that prior work. This type of information is hard to come by with current search technology. Neither the author's abstract, nor raw citation counts help users in assessing the relation between articles.

Fig. 1 shows a hypothetical search tool which displays differences and similarities between a target paper (here: *Pereira et al., 1993*) and the papers that it cites and that cite it. *Contrastive* links are shown in grey – links to rival papers and papers the current paper contrasts itself to. *Continuative* links are shown in black – links to papers that *use* the methodology of the current paper. Fig. 1 also displays the most characteristic textual sentence about each citation. For instance, we can see which aspect of *Hindle (1990)* our example paper criticises, and in which way the example paper's work was used by *Dagan et al. (1994)*.

Note that not even the CiteSeer text snippet can fulfil the relation search need: it is always centered around the physical location of the citations, but the context is often not informative enough for the searcher to infer the relation. In fact, studies from our annotated corpus (Teufel, 1999) show that 69% of the 600 sentences stating contrast with other work and 21% of the 246 sentences stating research continuation with other work do not contain the corresponding citation; the citation is found in preceding sentences (which means that the sentence expressing the contrast or continuation is outside the CiteSeer snippet). A more sophisticated, discourse-aware citation indexer which finds these sentences and associates them with the citation would add considerable value to the researcher's bibliographic search (Ritchie et al., 2006b).

Our annotation scheme for citations is based on empirical work in content citation analysis. It is designed for information retrieval applications such as improved citation indexing and better bibliometric measures (Teufel et al., 2006). Its 12 categories mark relationships with other works. Each citation is labelled with exactly one category. The following top-level four-way distinction applies:

- Explicit statement of weakness
- Contrast or comparison with other work (4 categories)
- Agreement/usage/compatibility with other work (6 categories), and
- A neutral category.

In this paper, we show that the scheme can be reliably annotated by independent coders. We also report results of a supervised machine learning experiment which replicates the human annotation.

2 An annotation scheme for citations

Our scheme (given in Fig. 2) is adapted from Spiegel-Rüsing's (1977) after an analysis of a corpus of scientific articles in computational linguistics. We avoid sociologically orientated distinctions ("paying homage to pioneers"), as they can be difficult to operationalise without deep knowledge of the field and its participants (Swales, 1986). Our redefinition of the categories aims at reliably annotation; at the same time, the categories should be informative enough for the document management application sketched in the introduction.

[1] These tools automatically citation-index all scientific articles reached by a web-crawler, making them available to searchers via authors or keywords in the title, and displaying the citation in context of a text snippet.

Category	Description
Weak	Weakness of cited approach
CoCoGM	Contrast/Comparison in Goals or Methods(neutral)
CoCo-	Author's work is stated to be superior to cited work
CoCoR0	Contrast/Comparison in Results (neutral)
CoCoXY	Contrast between 2 cited methods
PBas	Author uses cited work as basis or starting point
PUse	Author uses tools/algorithms/data/definitions
PModi	Author adapts or modifies tools/algorithms/data
PMot	This citation is positive about approach used or problem addressed (used to motivate work in current paper)
PSim	Author's work and cited work are similar
PSup	Author's work and cited work are compatible/provide support for each other
Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted citation function

Figure 2: Annotation scheme for citation function.

Our categories are as follows: One category (Weak) is reserved for weakness of previous research, if it is addressed by the authors. The next four categories describe comparisons or contrasts between own and other work. The difference between them concerns whether the contrast is between methods employed or goals (CoCoGM), or results, and in the case of results, a difference is made between the cited results being worse than the current work (CoCo-), or comparable or better results (CoCoR0). As well as considering differences between the current work and other work, we also mark citations if they are explicitly compared and contrasted with *other* work (i.e. not the work in the current paper). This is expressed in category CoCoXY. While this is not typically annotated in the literature, we expect a potential practical benefit of this category for our application, particularly in searches for differences and rival approaches.

The next set of categories we propose concerns positive sentiment expressed towards a citation, or a statement that the other work is actively used in the current work (which we consider the ultimate praise). We mark statements of use of data and methods of the cited work, differentiating unchanged use (PUse) from use with adaptations (PModi). Work which is stated as the explicit starting point or intellectual ancestry is marked with our category PBas. If a claim in the literature is used to strengthen the authors' argument, or vice versa, we assign the category PSup. We also mark similarity of (an aspect of) the approach to the cited work (PSim), and motivation of approach used or problem addressed (PMot).

Our twelfth category, Neut, bundles truly neutral descriptions of cited work with those cases where the textual evidence for a citation function was not enough to warrant annotation of that category, and all other functions for which our scheme did not provide a specific category.

Citation function is hard to annotate because it in principle requires interpretation of author intentions (what could the author's intention have been in choosing a certain citation?). One of our most fundamental principles is thus to only mark explicitly signalled citation functions. Our guidelines explicitly state that a general linguistic phrase such as "better" or "used by us" must be present; this increases the objectivity of defining citation function. Annotators must be able to point to textual evidence for assigning a particular function (and are asked to type the source of this evidence into the annotation tool for each citation). Categories are defined in terms of certain objective types of statements (e.g., there are 7 cases for PMot, e.g. "Citation claims that or gives reasons for why problem Y is hard"). Annotators can use general text interpretation principles when assigning the categories (such as anaphora resolution and parallel constructions), but are not allowed to use in-depth knowledge of the field or of the authors.

Guidelines (25 pages, $\sim$ 150 rules) describe the categories with examples, provide a decision tree and give decision aids in systematically ambiguous cases. Nevertheless, subjective judgement of the annotators is still necessary to assign a single tag in an unseen context, because of the many difficult cases for annotation. Some of these concern the fact that authors do not always state their purpose clearly. For instance, several earlier studies found that negational citations are rare (Moravcsik and Murugesan, 1975; Spiegel-Rüsing, 1977); MacRoberts and MacRoberts (1984) argue that the reason for this is that they are potentially politically dangerous. In our data we found ample evidence of the "meekness" effect. Other difficulties concern the distinction of the usage of a method from statements of similarity between a method and the own method (i.e., the choice between categories PSim and PUse). This happens in cases where authors do not want to admit (or stress)

that they are using somebody else's method. Another difficult distinction concerns the judgement of whether the authors *continue* somebody's research (i.e., consider their research as intellectual ancestry, i.e. `PBas`), or whether they simply *use* the work (`PUse`).

The unit of annotation is a) the full citation (as recognised by our automatic citation processor on our corpus), and b) names of authors of cited papers anywhere in running text outside of a formal citation context (i.e., without date). These latter are marked up, slightly unusually in comparison to other citation indexers, because we believe they function as important referents comparable in importance to formal citations.[2] In principle, there are many other linguistic expressions by which the authors could refer to other people's work: pronouns, abbreviations such as "Mueller and Sag (1990), henceforth M & S", and names of approaches or theories which are associated with particular authors. The fact that in these contexts citation function cannot be annotated (because it is not technically feasible to recognise them well enough) sometimes causes problems with context dependencies.

While there are unambiguous example cases where the citation function can be decided on the basis of the sentence alone, this is not always the case. In example (1) above the citation and the weakness occur in the same sentence, but it is more likely that a cited approach is neutrally described (often several sentences long), with the evaluative statement following much later (at the end of the textual segment about this citation). Nevertheless, the function must be marked on the nearest appropriate annotation unit (citation or author name). Our rules decree that context is in most cases constrained to the paragraph boundary. In rare cases, paper-wide information is required (e.g., for `PMot`, we need to know that a praised approach is used by the authors, information which may not be local in the paragraph). Annotators are thus asked to skim-read the paper before annotation.

One possible view on this annotation scheme could consider the first two sets of categories as "negative" and the third set of categories "positive", in the sense of Pang et al. (2002) and Turney (2002). Authors need to make a point (namely, that they have contributed something which is better or at least new (Myers, 1992)), and they thus have a stance towards their citations. But although there is a sentiment aspect to the interpretation of citations, this is not the whole story. Many of our "positive" categories are more concerned with different ways in which the cited work is useful to the current work (which aspect of it is used, e.g., just a definition or the entire solution?), and many of the contrastive statements have no negative connotation at all and simply state a (value-free) difference between approaches. However, if one looks at the distribution of positive and negative adjectives around citations, it is clear that there is a non-trivial connection between our task and sentiment classification.

The data we use comes from our corpus of 360 conference articles in computational linguistics, drawn from the Computation and Language E-Print Archive (http://xxx.lanl.gov/cmp-lg). The articles are transformed into XML format; headlines, titles, authors and reference list items are automatically marked up. Reference lists are parsed using regular patterns, and cited authors' names are identified. Our citation parser then finds citations and author names in running text and marks them up. Ritchie et al. (2006a) reports high accuracy for this task (94% of citations recognised, provided the reference list was error-free). On average, our papers contain 26.8 citation instances in running text[3]. For human annotation, we use our own annotation tool based on XML/XSLT technology, which allows us to use a web browser to interactively assign one of the 12 tags (presented as a pull-down list) to each citation.

We measure inter-annotator agreement between three annotators (the three authors), who independently annotated 26 articles with the scheme (containing a total of 120,000 running words and 548 citations), using the written guidelines. The guidelines were developed on a different set of articles from the ones used for annotation.

Inter-annotator agreement was Kappa=.72 (n=12;N=548;k=3)[4]. This is quite high, considering the number of categories and the difficulties

[2] Our citation processor can recognise these after parsing the citation list.

[3] As opposed to reference list items, which are fewer.

[4] Following Carletta (1996), we measure agreement in Kappa, which follows the formula $K = \frac{P(A)-P(E)}{1-P(E)}$ where P(A) is observed, and P(E) expected agreement. Kappa ranges between -1 and 1. K=0 means agreement is only as expected by chance. Generally, Kappas of 0.8 are considered stable, and Kappas of .69 as marginally stable, according to the strictest scheme applied in the field.

(e.g., non-local dependencies) of the task. The relative frequency of each category observed in the annotation is listed in Fig. 3. As expected, the distribution is very skewed, with more than 60% of the citations of category `Neut`.[5] What is interesting is the relatively high frequency of usage categories (`PUse, PModi, PBas`) with a total of 18.9%. There is a relatively low frequency of clearly negative citations (`Weak, CoCo-`, total of 4.1%), whereas the neutral–contrastive categories (`CoCoR0, CoCoXY, CoCoGM`) are slightly more frequent at 7.6%. This is in concordance with earlier annotation experiments (Moravcsik and Murugesan, 1975; Spiegel-Rüsing, 1977).

3 Features for automatic recognition of citation function

This section summarises the features we use for machine learning citation function. Some of these features were previously found useful for a different application, namely Argumentative Zoning (Teufel, 1999; Teufel and Moens, 2002), some are specific to citation classification.

3.1 Cue phrases

Myers (1992) calls meta-discourse the set of expressions that talk *about* the act of presenting research in a paper, rather than the research itself (which is called object-level discourse). For instance, Swales (1990) names phrases such as "*to our knowledge, no...*" or "*As far as we aware*" as meta-discourse associated with a gap in the current literature. Strings such as these have been used in extractive summarisation successfully ever since Paice's (1981) work.

We model meta-discourse (cue phrases) and treat it differently from object-level discourse. There are two different mechanisms: A finite grammar over strings with a placeholder mechanism for POS and for sets of similar words which can be substituted into a string-based cue phrase (Teufel, 1999). The grammar corresponds to 1762 cue phrases. It was developed on 80 papers which are different to the papers used for our experiments here.

The other mechanism is a POS-based recogniser of agents and a recogniser for specific actions these agents perform. Two main agent types (the authors of the paper, and everybody else) are modelled by 185 patterns. For instance, in a paragraph describing related work, we expect to find references to other people in subject position more often than in the section detailing the authors' own methods, whereas in the background section, we often find general subjects such as "*researchers in computational linguistics"* or *"in the literature"*. For each sentence to be classified, its grammatical subject is determined by POS patterns and, if possible, classified as one of these agent types. We also use the observation that in sentences without meta-discourse, one can assume that agenthood has not changed.

20 different action types model the main verbs involved in meta-discourse. For instance, there is a set of verbs that is often used when the overall scientific goal of a paper is defined. These are the verbs of presentation, such as "*propose, present, report*" and "*suggest*"; in the corpus we found other verbs in this function, but with a lower frequency, namely "*describe, discuss, give, introduce, put forward, show, sketch, state*" and "*talk about*". There are also specialised verb clusters which co-occur with `PBas` sentences, e.g., the cluster of continuation of ideas (eg. *"adopt, agree with, base, be based on, be derived from, be originated in, be inspired by, borrow, build on,... "*). On the other hand, the semantics of verbs in `Weak` sentences is often concerned with failing (of other researchers' approaches), and often contain verbs such as *"abound, aggravate, arise, be cursed, be incapable of, be forced to, be limited to, ... "*.

We use 20 manually acquired verb clusters. Negation is recognised, but too rare to define its own clusters: out of the $20 \times 2 = 40$ theoretically possible verb clusters, only 27 were observed in our development corpus. We have recently automated the process of verb–object pair acquisition from corpora for two types of cue phrases (Abdalla and Teufel, 2006) and are planning on expanding this work to other cue phrases.

3.2 Cues Identified by annotators

During the annotator training phase, the annotators were encouraged to type in the meta-description cue phrases that justify their choice of category. We went through this list by hand and extracted 892 cue phrases (around 75 per category). The files these cues came from were not part of the test corpus. We included 12 features

[5] Spiegel-Rüsing found that out of 2309 citations she examined, 80% substantiated statements.

Neut	PUse	CoCoGM	PSim	Weak	PMot	CoCoR0	PBas	CoCoXY	CoCo-	PModi	PSup
62.7%	15.8%	3.9%	3.8%	3.1%	2.2%	0.8%	1.5%	2.9%	1.0%	1.6%	1.1%

Figure 3: Distribution of citation categories

that recorded the presence of cues that our annotators associated with a particular class.

3.3 Other features

There are other features which we use for this task. We know from Teufel and Moens (2002) that verb tense and voice should be useful for recognizing statements of previous work, future work and work performed in the paper. We also recognise modality (whether or not a main verb is modified by an auxiliary, and which auxiliary it is).

The overall location of a sentence containing a reference should be relevant. We observe that more `PMot` categories appear towards the beginning of the paper, as do `Weak` citations, whereas comparative results (`CoCoR0`, `CoCoR-`) appear towards the end of articles. More fine-grained location features, such as the location within the paragraph and the section, have also been implemented.

The fact that a citation points to own previous work can be recognised, as we know who the paper authors are. As we have access to the information in the reference list, we also know the last names of *all* cited authors (even in the case where an *et al.* statement in running text obscures the later-occurring authors). With self-citations, one might assume that the probability of re-use of material from previous own work should be higher, and the tendency to criticise lower.

4 Results

	Weakness	Positive	Contrast	Neutral
P	.80	.75	.77	.81
R	.49	.65	.52	.90
F	.61	.70	.62	.86
Percentage Accuracy				0.79
Kappa (n=12; N=2829; k=2)				0.59
Macro-F				0.68

Figure 5: Summary of results (10-fold cross-validation; IBk algorithm; k=3): Top level classes.

Our evaluation corpus for citation analysis consists of 116 articles (randomly drawn from the part of our corpus which was not used for human annotation, for guideline development or cue phrase development). The 116 articles contain 2829 citation instances. Each citation instance was manually tagged as one of {`Weak`, `CoCoGM`, `CoCo-`, `CoCoR0`, `CoCoXY`, `PBas`, `PUse`, `PModi`, `PMot`, `PSim`, `PSup`, `Neut`}. The papers are then automatically processed: POS-tagged, self-citations are detected by overlap of citing and cited authors, and all other features are identified before the machine learning is applied.

	Weakness	Positive	Neutral
P	.77	.75	.85
R	.42	.65	.92
F	.54	.70	.89
Percentage Accuracy			0.83
Kappa (n=12; N=2829; k=2)			0.58
Macro-F			0.71

Figure 6: Summary of results (10-fold cross-validation; IBk algorithm; k=3): Sentiment Analysis.

The 10-fold cross-validation results for citation classification are given in Figure 4, comparing the system to one of the annotators. Results are given in three overall measures: Kappa, percentage accuracy, and Macro-F (following Lewis (1991)). Macro-F is the mean of the F-measures of all twelve categories. We use Macro-F and Kappa because we want to measure success particularly on the rare categories, and because Micro-averaging techniques like percentage accuracy tend to overestimate the contribution of frequent categories in heavily skewed distributions like ours[6].

In the case of Macro-F, each category is treated as one unit, independent of the number of items contained in it. Therefore, the classification success of the individual items in rare categories is given more importance than classification success of frequent category items. However, one should keep in mind that numerical values in macro-averaging are generally lower (Yang and Liu, 1999), due to fewer training cases for the rare categories. Kappa has the additional advantage over Macro-F that it filters out random agreement (random use, but following the observed distribu-

[6]This situation has parallels in information retrieval, where precision and recall are used because accuracy overestimates the performance on irrelevant items.

	Weak	CoCoGM	CoCoR0	CoCo-	CoCoXY	PBas	PUse	PModi	PMot	PSim	PSup	Neut
P	.78	.81	.77	.56	.72	.76	.66	.60	.75	.68	.83	.80
R	.49	.52	.46	.19	.54	.46	.61	.27	.64	.38	.32	.92
F	.60	.64	.57	.28	.62	.58	.63	.37	.69	.48	.47	.86

Percentage Accuracy	0.77
Kappa (n=12; N=2829; k=2)	0.57
Macro-F	0.57

Figure 4: Summary of Citation Analysis results (10-fold cross-validation; IBk algorithm; k=3).

tion of categories).

For our task, memory-based learning outperformed other models. The reported results use the IBk algorithm with $k = 3$ (we used the Weka machine learning toolkit (Witten and Frank, 2005) for our experiments). Fig. 7 provides a few examples from one file in the corpus, along with the gold standard citation class, the machine prediction, and a comment.

Kappa is even higher for the top level distinction. We collapsed the obvious similar categories (all `P` categories into one category, and all `CoCo` categories into another) to give four top level categories (`Weak, Positive, Contrast, Neutral`; results in Fig. 5). Precision for all the categories is above 0.75, and K=0.59. For contrast, the human agreement for this situation was K=0.76 (n=3,N=548,k=3).

In a different experiment, we grouped the categories as follows, in an attempt to perform sentiment analysis over the classifications:

Old Categories	New Category
Weak, CoCo-	Negative
PMot, PUse, PBas, PModi, PSim, PSup	Positive
CoCoGM, CoCoR0, CoCoXY, Neut	Neutral

Thus negative contrasts and weaknesses are grouped into `Negative`, while neutral contrasts are grouped into `Neutral`. All positive classes are conflated into `Positive`.

Results show that this grouping raises results to a smaller degree than the top-level distinction did (to K=.58). For contrast, the human agreement for these collapsed categories was K=.75 (n=3,N=548,k=3).

5 Conclusion

We have presented a new task: annotation of citation function in scientific text, a phenomenon which we believe to be closely related to the overall discourse structure of scientific articles. Our annotation scheme concentrates on weaknesses of other work, and on similarities and contrast between work and usage of other work. In this paper, we present machine learning experiments for replicating the human annotation (which is reliable at K=.72). The automatic result reached K=.57 (acc=.77) for the full annotation scheme; rising to Kappa=.58 (acc=.83) for a three-way classification (Weak, Positive, Neutral).

We are currently performing an experiment to see if citation processing can increase performance in a large-scale, real-world information retrieval task, by creating a test collection of researchers' queries and relevant documents for these (Ritchie et al., 2006a).

6 Acknowledgements

This work was funded by the EPSRC projects CITRAZ (GR/S27832/01, "Rhetorical Citation Maps and Domain-independent Argumentative Zoning") and SCIBORG (EP/C010035/1, "Extracting the Science from Scientific Publications").

References

Rashid M. Abdalla and Simone Teufel. 2006. A bootstrapping approach to unsupervised detection of cue phrase variants. In *Proc. of ACL/COLING-06*.

Susan Bonzi. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *JASIS*, 33(4):208–216.

Christine L. Borgman, editor. 1990. *Scholarly Communication and Bibliometrics*. Sage Publications, CA.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Daryl E. Chubin and S. D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4):423–441.

Eugene Garfield. 1979. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. J. Wiley, New York, NY.

Eugene Garfield. 1990. How ISI selects journals for coverage: Quantitative and Qualitative considerations. Current Contents, May 28.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: An automatic citation indexing system. In *Proc. of the Third ACM Conference on Digital Libraries*, pages 89–98.

Context	Human	Machine	Comment
We have compared four complete and three partial data representation formats for the baseNP recognition task presented in **Ramshaw and Marcus (1995)**.	PUse	PUse	Cues can be weak: "*for... task... presented in*"
In the version of the algorithm that we have used, IB1-IG, the distances between feature representations are computed as the weighted sum of distances between individual features **(Bosch 1998)**.	Neut	PUse	Human decided citation was for detail in used package, not directly used by paper.
We have used the baseNP data presented in **Ramshaw and Marcus (1995)**.	PUse	PUse	Straightforward case
We will follow **Argamon et al. (1998)** and use a combination of the precision and recall rates: F=(2*precision*recall)/(precision+recall).	PSim	PUse	Human decided F-measure was not attributable to citation. Hence similarity rather than usage.
This algorithm standardly uses the single training item closest to the test i.e. However **Daelemans et al. (1999)** report that for baseNP recognition better results can be obtained by making the algorithm consider the classification values of the three closest training items.	Neut	PUse	Shallow processing by Machine means that it is mislead by the strong cue in preceding sentence.
They are better than the results for section 15 because more training data was used in these experiments. Again the best result was obtained with IOB1 (F=92.37) which is an improvement of the best reported F-rate for this data set ((**Ramshaw and Marcus 1995)** (F=92.03).	CoCo-	PUse	Machine is misled by strong cue for usage in preceding sentence.

Figure 7: Examples of classifications by the machine learner.

T.L. Hodges. 1972. *Citation Indexing: Its Potential for Bibliographical Control*. Ph.D. thesis, University of California at Berkeley.

David D. Lewis. 1991. Evaluating text categorisation. In *Speech and Natural Language: Proceedings of the ARPA Workshop of Human Language Technology*.

Terttu Luukkonen. 1992. Is scientists' publishing behaviour reward-seeking? *Scientometrics*, 24:297–319.

Michael H. MacRoberts and Barbara R. MacRoberts. 1984. The negational reference: Or the art of dissembling. *Social Studies of Science*, 14:91–94.

Michael J. Moravcsik and Poovanalingan Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5:88–91.

Greg Myers. 1992. In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics*, 17(4).

John O'Connor. 1982. Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management*, 18(3):125–131.

Chris D. Paice. 1981. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In R. Oddy, S. Robertson, C. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*. Butterworth, London, UK.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP-02*.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006a. Creating a test collection for citation-based IR experiments. In *Proc. of HLT/NAACL 2006*, New York, US.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006b. How to find better index terms through citations. In *Proc. of ACL/COLING workshop "Can Computational Linguistics improve IR"*.

Simon Buckingham Shum. 1998. Evolving the web for scientific knowledge: First steps towards an "HCI knowledge web". *Interfaces, British HCI Group Magazine*, 39.

Henry Small. 1982. Citation context analysis. In P. Dervin and M. J. Voigt, editors, *Progress in Communication Sciences 3*, pages 287–310. Ablex, Norwood, N.J.

Ina Spiegel-Rüsing. 1977. Bibliometric and content analysis. *Social Studies of Science*, 7:97–113.

John Swales. 1986. Citation analysis and discourse analysis. *Applied Linguistics*, 7(1):39–56.

John Swales, 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*, pages 110–176. Cambridge University Press, Cambridge, UK.

Simone Teufel and Marc Moens. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–446.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. An annotation scheme for citation function. In *Proc. of SIGDial-06*.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, UK.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL-02*.

Melvin Weinstock. 1971. Citation indexes. In *Encyclopedia of Library and Information Science*, volume 5. Dekker, New York, NY.

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

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proc. of SIGIR-99*.

John M. Ziman. 1968. *Public Knowledge: An Essay Concerning the Social Dimensions of Science*. Cambridge University Press, Cambridge, UK.