Argumentative Zoning for improved citation indexing

Simone Teufel
Computer Laboratory
University of Cambridge
JJ Thomson Avenue, Cambridge CB3 0FD, UK
Simone.Teufel@cam.ac.uk

Abstract

We address the problem of automatically classifying academic citations in scientific articles according to author affect. There are many ways how a citation might fit into the overall argumentation of the article: as part of the solution, as rival approach or as flawed approach that justifies the current research. Our motivation for this work is to improve citation indexing. The method we use for this task is machine learning from indicators of affect (such as "we follow X in assuming that...", or "in contrast to Y, our system solves this problem") and of presentation of ownership of ideas (such as "We present a new method for...", or "They claim that..."). Some of these features are borrowed from Argumentative Zoning (Teufel & Moens 2002), a technique for determining the rhetorical status of each sentence in a scientific article. These features include the type of subject of the sentence, the citation type, the semantic class of main verb, and a list of indicator phrases. Evaluation will be both intrinsic and extrinsic, involving the measurement of human agreement on the task and a comparison of human and automatic evaluation, as well as a comparison of task-performance with our system versus task performance with a standard citation indexer (CiteSeer, (Lawrence, Giles, & Bollacker 1999)).

Citation Indexing and Citation Maps

Automatic indexing, as exemplified by the highly successful tool CiteSeer (Giles, Bollacker, & Lawrence 1998), has become the method of choice for literature searches; as a result, CiteSeer receives more than 8000 hits a day. CiteSeer automatically citation-indexes all scientific articles reached by a web-crawler, making them available to searchers via authors or keywords in the title.

However, keywords are not everything in literature searches. Shum (1998) states that researchers, particularly experienced researchers, are often interested in relations between articles. They need 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. And even though CiteSeer shows a text snippet around the physical location for searchers to peruse, there is no guarantee that the text snippet provides enough information for the searcher to infer the relation.

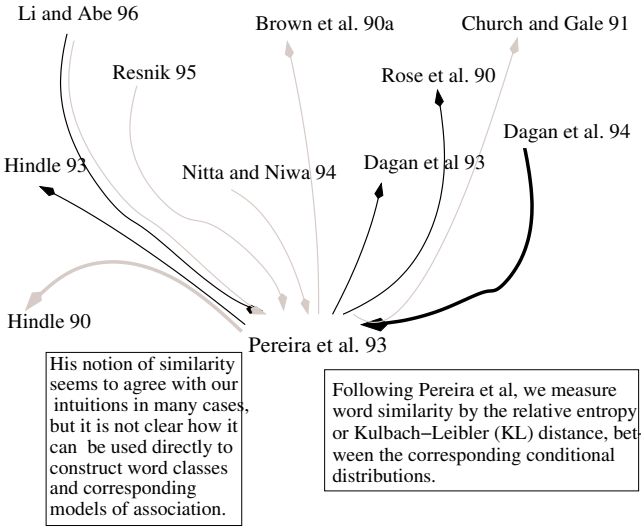

Figure 1: A rhetorical citation map

Being able to interpret the rhetorical status of a citation at a glance would add considerable value to citation indexes, as shown in Fig. 1. Here differences and similarities are shown between the example paper (*Pereira et al., 1993*) and the papers it cites, as well as the papers that cite it – within the universe of our smallish corpus of scientific papers. We distinguish *contrastive* links (shown in grey) – links to rival papers and papers the current paper contrasts itself to – and *continuative* links (shown in black) – links to papers that are taken as starting point of the current research, or as part of the methodology of the current paper. In the citation map, the most important textual sentence about each citation can be displayed; these sentences are extracted from the original text. For instance, the map tells us which aspect of *Hindle (1990)* the *Pereira et al.* paper criticises, and in which way *Pereira et al.*'s work was used by *Dagan et al.* (1994).

In a larger context (ie. with thousands of citations automatically citation-indexed), we would be able to trace the

[Pereira et al, 1993] $\rightarrow_{basis}$ [Dagan et al, 1993]	**155** The data for this test was built from the training data for the previous one in the following way, based on a suggestion by Dagan et al. (1993).	
[Pereira et al, 1993] $\rightarrow_{contrast}$ [Resnik, 1992]	**11** While it may be worthwhile to base such a model on preexisting sense classes (Resnik 1992), in the work described here we look at how to derive the classes directly from distributional data.	
[Resnik, 1995] $\rightarrow_{contrast}$ [Pereira et al, 1993]	**0** Word groupings useful for language processing tasks are increasingly available [...] (e.g. Bensch and Savitch (1992), [...], Pereira et al. (1993), Schuetze (1993).	**1** However, for many tasks, one is interested in relationships among word senses, not words.

Figure 2: Some of Pereira et al. (1993)'s citation relations in our corpus

citation relations of our example paper across time; Fig. 2 shows part of such information (with sentence numbers indicating where in the text these sentences were taken from).

Simple citation parsing and displaying of sentences containing citations is not enough to achieve this type of output. CiteSeer makes the simplifying assumption that the most important information about a citation is always local to the physical citation. This assumption does not hold. In the annotated corpus from Teufel and Moens (2002), where sentences are marked up according to rhetorical context, we found that 69% of the 600 evaluative CONTRAST sentences and 21% of the 246 BASIS sentences do not contain the corresponding citation; the citation is found in preceding sentences instead. Therefore, CiteSeer will miss to display the evaluative statement in many cases. Nanba and Okumura (1999) present an automatic citation indexer which, like ours, automatically classifies contexts (in their case, into "positive" and "negative" contexts). However, they display a large context of around 3 sentences per citation, assuming that the important sentence expressing author affect is in this area. In our approach, we aim to find the single sentence containing the evaluative statement that connects two papers, even if that sentence is textually removed from the citation. Therefore, our approach makes maximally short and informative descriptions possible. We relies on corpus-based discourse analysis to find this sentence.

The task of building citation maps can be formulated as a statistical classification problem. For each evaluative statement identified by Argumentative Zoning, we determine a set of potential candidate citation identifiers nearby, and use machine learning to associate the correct candidate identifier with the evaluative statement. The output of the classifier is a citation and a corresponding evaluative statement (a sentence), which can be displayed in the citation map.

Argumentative Zoning and Author Affect

Scientific writing is supposed to be objective and affect-free, but it is not. In fact, scientific texts are often so full of subjective statements, fixed phrases and hedges that even rather shallow techniques can exploit this fact to improve text understanding.

One example of such a technique is Argumentative Zoning (AZ; Teufel & Moens, 2002) , a shallow method of discourse analysis which automatically determines the rhetorical status of each sentence in a text as one of the seven rhetorical roles defined in Fig. 3 (examples from our corpus in Fig. 4, with CMP_LG identifiers (CMP_ LG 1994)).

AIM	Specific research goal of the current paper
TEXTUAL	Statements about section structure
OWN	(Neutral) description of own work presented in current paper
BACKGROUND	Generally accepted scientific background
CONTRAST	Comparison with or contrast to other work
BASIS	Statements of agreement with other work or continuation of other work
OTHER	(Neutral) description of other researchers' work

Figure 3: Argumentative Zoning: Categories

AIM	*We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts* (9408011).
TEXTUAL	*This section presents a morphographemic model which handles error detection in non-linear strings* (9504024).
OWN	*Our model associates phrases with relation graphs (9408014).*
BACKGROUND	*Semitic is known amongst computational linguists for its highly inflexional morphology* (9504024).
CONTRAST	*However, Solomonoff does not give a concrete search algorithm and only makes suggestions as to its nature* (9504034).
BASIS	*We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988)* (9504007).
OTHER	*The semidirectional Lambek calculus (henceforth SDL) is a variant of J. Lambek's original calculus of syntactic types (Lambek 1958)* (9605016).

Figure 4: Argumentative Zoning: Examples

The categories CONTRAST and BASIS are directly relevant to the citation classification work described here. These two types of sentences are also the ones which are particularly concerned with affect, as they correspond roughly to positive and negative descriptions of other researcher's work. Of course, "positive" and "negative" affect are over-

simplifications of much finer classification schemes developed in the field of Content Citation Analysis. This work has concentrated on manual annotation of the function of each citation (for an overview cf. (Weinstock 1971)). While we feel that our two categories are a workable approximation of these schemes for automation purposes, we remain interested in the fuller annotation schemes for the longer-term future.

1. AbsLoc	Position of sentence; 10 segments A-J
2. SectStruct	Relative and absolute position of sentence within section (e.g., first sentence in section or Last Third; 7 values
3. ParaStruct	Relative position of sentence within a paragraph;Initial, Medial, Final
4. HeadLine	Type of headline of current section; 16 classes
5. SentLength	Is the sentence longer than a certain threshold?
6. TitleContent	Does the sentence contain words also occurring in the title or headlines?
7. TF*IDF Content	Does the sentence contain "significant terms" as determined by the *TF*IDF* measure?
8. VerbVoice	Voice (of first finite verb in sentence); Active or Passive or NoVerb
9. VerbTense	Tense (of first finite verb in sentence); 9 tenses or NoVerb
10. VerbModal	Is the first finite verb modified by modal auxiliary?
11. Cit	Does the sentence contain a citation or the name of an author contained in the reference list? If it contains a citation, is it a self citation? Whereabouts in the sentence does the citation occur? ({Citation (self), Citation (other), Author Name, or None} X {Beginning, Middle, End})
12. History	Most probable previous category; 7 Target Categories + "BEGIN"
13. Formulaic	Type of formulaic expression occurring in sentence; 18 Types + 9 Agent Types or None
14. Agent	Type of Agent; 9 Agent Types or None
15. Action	Type of Action, with or without Negation; 27 Action Types or None

Figure 5: Features used for Argumentative Zoning

AZ is based on machine learning with the Naive Bayes classifier, as in the Kupiec, Pedersen, & Chen (1995) approach to statistical sentence extraction. 15 features are used (cf. the overview in Figure 5), some of which are borrowed from the sentence extraction literature (such as location of a sentence in the article, or the sum of the relative term frequencies of the content words contained in it), and some of which are new and linguistically more interesting (such as the attribution-type of the subject). For instance, in order to find out if a sentence is part of the BACKGROUND section or the OWN section, knowing that the subject of the sentence is "our system" might bias one towards the OWN section. Feature determination is shallow in that it requires only POS-tagging. The material used to train the system were 80 papers (around 12,000 sentences) which were manually annotated, with reasonable inter- and intra-annotator agreement (Teufel, Carletta, & Moens 1999).

The original application of AZ was summarisation: Extractive summaries can be formed by choosing particularly important labels (e.g. sc Aim, CONTRAST and BASIS) and by selecting those sentences which have the highest probabilistic score for that given label. The experiment in (Teufel 2001), where the task used to evaluate the quality of abstracts was to list related articles and their relationship to the current paper, indicates that AZ information could be very useful in the short run to improve citation indexes. Subjects with AZ-extracts were able to perform this task almost as well as a control group given the full papers.

Meta-discourse

One set of features particularly interesting for citation classification are the so-called meta-discourse features. As meta-discourse we understand here, in the sense of Myers (1992), the set of expressions that talk *about* the act of presenting research in a paper, rather than the research itself. Swales (1990) found that the argumentation of the paper is rather prototypical; it might start by convincing us that the research done in the paper is hard or difficult, and that there is a gap in the current literature. This gap, for instance, is often indicated by a phrase such as "*to our knowledge, no...*" or "*As far as we aware*". The Formulaic feature collects 1762 such phrases and their variations.

The feature Agent models the succession of grammatical subjects in meta-discourse, which often signal who the ideas in a given paragraph are attributed to. 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*". There is also a strong segmental aspect to the phenomenon of attribution of authorship: in sentences without meta-discourse, one assumes that the same sets of players (the authors, their rivals, or general researchers in the area) are still active. These assumptions are modelled in the Agent feature, which maps every sentence to 10 different classes of agents.

From a viewpoint of lexical semantics, it is interesting to look at the main verbs involved in meta-discourse. This is expressed in the Action feature. 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 specialised verb clusters which co-occur with BASIS sentences,

adopt, agree with, base, be based on, be derived from, be originated in, be inspired by, borrow, build on, follow, originate from, originate in, side with

Figure 6: Verbs of continuation

adapt, adjust, augment, combine, change, decrease, elaborate on, expand, extend, derive, incorporate, increase, manipulate, modify, optimize, refine, render, replace, revise, substitute, tailor, upgrade

Figure 7: Verbs of change

abound, aggravate, arise, be cursed, be incapable of, be forced to, be limited to, be problematic, be restricted to, be troubled, be unable to, contradict, damage, degrade, degenerate, fail, fall prey, fall short, force oneself, force, hinder, impair, impede, inhibit, lack, misclassify, misjudge, mistake, misuse, neglect, obscure, overestimate, overfit, overgeneralize, overgenerate, overlook, pose, plague, preclude, prevent, resort to, restrain, run into problems, settle for, spoil, suffer from, threaten, thwart, underestimate, undergenerate, violate, waste, worsen

Figure 8: Verbs of failure

be different from, be distinct from, conflict, contrast, clash, differ from, distinguish oneself, differentiate, disagree, disagreeing, dissent, oppose

Figure 9: Verbs of contrast

e.g. the verb semantics of continuation of ideas (cf. Fig 6) or of change (cf. Fig 7).

On the other hand, the semantics of verbs in CONTRAST sentences is often concerned with failing (of other researchers' ideas; cf. Fig 8) or contrast (cf. Fig 9).

Currently the verb clusters we use are manually collected; the feature `Action` maps them onto 20 features (in theory, there are twice as many as negation of the sentence is also taken into account and combined with these 20 groups – in practice only 27 of these Action Types occur in our corpus as negation is rare). In future work, we are interested in how to automate the process of verb cluster determination.

Human Annotation of Author Affect

In order to machine learn author affect, we have created a corpus of human annotated citation, starting from the annotations in Teufel & Moens (2002), where every sentence was associated with one of the seven categories. In that work, we used three annotators, written guidelines of 17 pages, and a formal training procedure of 7 hours. We measured intra- and inter-annotator agreement. Intra-annotator agreement, i.e. the similarity of the annotation of *one* annotator after a time period long enough for the annotator to have forgotten the original annotation, is important as it justifies the well-definedness of the semantic labels of an annotation

5 **(Hindle 1990) (Contrastive, +4)** proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen.
6 For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs.
7 This requires a reasonable definition of verb similarity and a similarity estimation method.
8 In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.
9 His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.

113 The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering **(Rose et al. 1990)**. **(Continuation, 0)**

Figure 10: Annotation examples

scheme. We concluded that our scheme was reasonably stable (K=.82, .81, .76) and reproducible (K=.71). The distribution of categories was very skewed: 67% OWN, 16% OTHER, 6% BACKGROUND, 5% CONTRAST, and 2% each for BASIS, AIM and TEXTUAL. Further analyses showed that AIM and TEXTUAL are categories the annotators were particularly good at determining, whereas BASIS and CONTRAST were relatively more difficult.

For the new project, a different type of annotation was necessary: for each evaluative statement (CONTRAST and BASIS), our annotators had to identify one or more (or zero) citations in the text. These citation could be either in the current sentence, or in sentences before or after the evaluative statement.

In Citation-indexing, more than one target classification must be determined:

- More than one citation can be associated with an evaluative statement
- The citation concerned can be in text before or after the evaluative statement
- The distance of the sentence expressing the evaluative statement from the citation must also be determined.

We have written an new set of guidelines, and currently have 1000 annotated sentence/citation pairs. We have not yet measured human agreement on the task.

Fig. 10 shows two example contexts from our sample paper with their citation annotation. In the first example, the evaluation of the citation *Hindle (1990)* is contrastive, and the evaluative statement is found 4 sentences after the sentence containing the citation. In the second example, the citation *Rose et al. (1990)* is evaluated as a continuation; the evaluative statement can be found in the same sentence as the physical citation.

During the annotation, we noticed typical patterns of citation and evaluation, which is illustrated in Fig. 11. The little square box signifies the citation itself; white background a neutral description of other work (OTHER); patterned background a neutral description of own work (OWN); dark grey

BASIS and light grey CONTRAST evaluation. a) and b) show normal cases where other work is cited, and either Contrast or Basis evaluation is expressed a few sentences later. In case c), the approach is identified (by citation) and criticised in the first sentence, only later is the approach described. This pattern is rarer than the corresponding pattern a). Pattern d) shows an approach which is cited but receives no evaluation. While citing without stating why one cites is against good writing advice, this pattern nevertheless occurred frequently in our corpus. Pattern e) is quite frequent for BASIS sentences: as they are often used to describe which other work forms part of the own methodology, BASIS sentences often occur embedded in OWN sentences. In these cases, the citation and the evaluation are typically present in the same sentence. In pattern f), CONTRAST sentences are embedded in OWN sentences, without repetition of the citation itself (which occurred higher up in the text). One situation in which we observed this pattern is when results of a rival approach are contrasted to the own results.

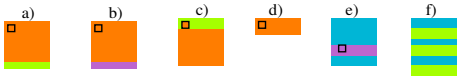

Figure 11: Patterns of citing and author stance statements

Features for Author Affect

The 15 features used for Argumentative Zoning were presented in Fig. 5 above, and the meta-discourse features were explained above. We will reuse some of the features used for Argumentative Zoning for the new citation classification task. In particular, the most successful features for Argumentative Zoning were (in descending order of usefulness): Absolute Sentence Location, Agent, Citations, Headlines, History, Formulaic, and Action. The other features only minimally improved results.

One of the AZ features, the history feature, models local context: it takes the category of the previous sentence as a feature, as there are often patterns of categories following each other. During testing, the category of the previous sentence is only probabilistically known, which is why a beam search is performed. We expect this feature to be equally important for citation relation learning.

Other methods from coreference resolution will be newly implemented for this project. The task of finding associations is loosely related to anaphora resolution. The differences between anaphora links and citation associations is that the latter appear less frequently in text, but seem to build links which are stronger, less ambiguous, and more global than anaphoric links. Constraining factors such as agreement information and WordNet relations, which prove very useful for anaphora resolution, are probably of less use for this task. We plan to borrow features from work such as Ge et al.'s (1998) : type of candidate identification, type of alternative candidates, type of citation (self citation or foreign citation), location in the document of the evaluative statement, direction of identification (forward in text, or backward) and saliency factors such as (estimated) grammatical function of identification phrase, verb and verb tense.

In addition to these features, we will also exploit regularities such as the ones described in Fig. 11 (patterns of citations and author stance).

Evaluation

Intrinsic evaluation of Argumentative Zoning was performed by measuring similarity of system annotation with human annotation, expressed in Kappa (Siegel & Castellan 1988) and Macro-F (Lewis 1991) (wrt. precision and recall of each of the seven categories). Our system showed an annotation accuracy of F=.50 (K=.45), beating a text classification baseline of F=.30 (K=.30), but remaining well under human performance (F=.69; K=.71). Extrinsic evaluation (Teufel 2001) showed that for a question-answering task which concentrated on relations between articles, AZ-enhanced sentence extracts were significantly more useful than any other short document representation (including authors' abstracts and traditional sentence extracts).

For the new project, two types of evaluation are planned. Again, the intrinsic evaluation will compare system annotation with human annotation. The extrinsic evaluation will evaluate the usefulness of citation maps in comparison with alternative document surrogates, using specific questions created for our development corpus. We will create 20 pairs of document + question pairs about related work in the CL domain. For instance, the question for an article which uses manual rules for genre identification might be "*Name another article with a different method for genre identification*", or "*Does this article use the same classification as Karlgren (1994)?*". We will ask experts to verify the correctness of our answers. We can then measure the accuracy and time required to answer these questions using citation maps, as opposed to using CiteSeer or a search engine such as Google.

Conclusion

The automatic detection of subjectivity and point-of-view is traditionally associated with genres such as novels and scientific writing (Wiebe 1994), and tasks such as sentiment classification have used reviews of financial services or movies as their texts (Pang, Lee, & Vaithyanathan 2002). We believe that scientific text also contains subjective content, and that this content can be determined and exploited for tasks such as summarisation and better citation indexing. Here, we have described a new task for citation-indexing that uses positive or negative spin on citations to guide users during their literature searches. This is an early project report; at this stage, we have created the training material and are currently adapting the features from Argumentative Zoning to this task.

References

CMP_LG. 1994. The Computation and Language E-Print Archive, `http://xxx.lanl.gov/cmp-lg`.

Ge, N.; Hale, J.; and Charniak, E. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

Giles, C. L.; Bollacker, K. D.; and Lawrence, S. 1998. Citeseer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, 89–98.

Kupiec, J.; Pedersen, J. O.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95)*, 68–73.

Lawrence, S.; Giles, C. L.; and Bollacker, K. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer* 32(6):67–71.

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

Myers, G. 1992. In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics* 17(4):295–313.

Nanba, H., and Okumura, M. 1999. Towards multi-paper summarization using reference information. In *Proceedings of IJCAI-99*, 926–931.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shum, S. B. 1998. Evolving the web for scientific knowledge: First steps towards an "HCI knowledge web". *Interfaces, British HCI Group Magazine* 39:16–21.

Siegel, S., and Castellan, N. J. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. Berkeley, CA: McGraw-Hill, 2nd edition.

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

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

Teufel, S.; Carletta, J.; and Moens, M. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 110–117.

Teufel, S. 2001. Task-based evaluation of summary quality: Describing relationships between scientific papers. In *Proceedings of NAACL-01 Workshop "Automatic Text Summarization"*.

Weinstock, M. 1971. Citation indexes. In *Encyclopedia of Library and Information Science*, volume 5. New York, NY: Dekker. 16–40.

Wiebe, J. 1994. Tracking point of view in narrative. *Computational Linguistics* 20(2):223–287.