# Whose idea was this, and why does it matter? Attributing scientific work to citations

**Advaith Siddharthan & Simone Teufel**
Natural Language and Information Processing Group
University of Cambridge Computer Laboratory
{as372,sht25}@cl.cam.ac.uk

## Abstract

Scientific papers revolve around citations, and for many discourse level tasks one needs to know whose work is being talked about at any point in the discourse. In this paper, we introduce the *scientific attribution* task, which links different linguistic expressions to citations. We discuss the suitability of different evaluation metrics and evaluate our classification approach to deciding attribution both intrinsically and in an extrinsic evaluation where information about scientific attribution is shown to improve performance on Argumentative Zoning, a rhetorical classification task.

## 1 Introduction

In the recent past, there has been a focus on information management from scientific literature. In the genetics domain, for instance, information extraction of genes and gene–protein interactions helps geneticists scan large amounts of information (e.g., as explored in the TREC Genomics track (Hersh et al., 2004)). Elsewhere, citation indexes (Garfield, 1979) provide bibliometric data about the frequency with which particular papers are cited. The success of citation indexers such as CiteSeer (Giles et al., 1998) and Google Scholar relies on the robust detection of formal citations in arbitrary text. In bibliographic information retrieval, anchor text, i.e., the context of a citation can be used to characterise (index) the cited paper using terms outside of that paper (Bradshaw, 2003); O'Connor (1982) presents an approach for identifying the area around citations where the text focuses on that citation. And automatic citation classification (Nanba and Okumura, 1999; Teufel et al., 2006) determines the function that a citation plays in the discourse.

For such information access and retrieval purposes, the relevance of a citation within a paper is often crucial. One can estimate how important a citation is by simply counting how often it occurs in the paper. But as Kim and Webber (2006) argue, this ignores many expressions in text which refer to the cited author's work but which are not as easy to recognise as citations. They address the resolution of instances of the third person personal pronoun *"they"* in astronomy papers: it can either refer to a citation or to some entities that are part of research within the paper (e.g., planets or galaxies). Several applications should profit in principle from detecting connections between referring expressions and citations. For instance, in citation function classification, the task is to find out if a citation is described as flawed or as useful. Consider:

> Most computational models of discourse are based primarily on an analysis of the intentions of the speakers **[Cohen and Perrault, 1979][Allen and Perrault, 1980][Grosz and Sidner, 1986]**$^{\text{WEAK}}$. The speaker will form intentions **based on** his goals and then act on these intentions, producing utterances. The hearer will then reconstruct a model of the speaker's intentions upon hearing the utterance. This approach has many strong points, but **does not provide a very satisfactory account** of the adherence to discourse conventions in dialogue.

The three citations above are described as flawed (detectable by *"does not provide a very satisfactory account"*), and thus receive the label WEAK. However, in order to detect this, one must first realise that *"this approach"* refers to

the three cited papers. A contrasting hypothesis could be that the citations are *used* (thus deserving the label PUse; the cue phrase *"based on"* might make us think so (as in the context *"our work is based on"*). This, however, can be ruled out if we know that *"the speaker"* is not referring to some aspect of the current paper.

## 2 The scientific attribution task

We define an attribution task where possible referents are members of the reference list (i.e., each cited paper), the Current-Paper, and a back-off category No-Specific-Paper for markables that are not attributable to any specific paper(s). Our markables are as follows: all definite descriptions (e.g., *"the hearer"*, and including demonstrative noun phrases such as *"these intentions"*), all "work" nouns[1], and all pronouns (possessive, personal and demonstrative); c.f., underlined strings in the above example. Our notion of attribution link encompasses two relations:

1. Anaphoric: The referents are entire research papers, or the papers' authors
2. Subpart: The referents are some component of an approach/argument/claim in the research paper

There are two tasks: attributing a linguistic expression to the right paper (including the current paper) – a task we call *scientific attribution* –, and deciding whether or not the expression is anaphoric to the entirety of the paper, or just to some subpart of it.

Kim and Webber (2006) solve the problem of distinguishing between these relations for one case. They decide whether the pronoun *"they"* anaphorically refers to the authors of a cited paper, or whether it refers to some entity that is discussed in (a subpart of) a paper (e.g., *"galaxies"*). In this paper, we tackle the other problem of scientific attribution.

We do not distinguish between the two types of links stated above, but only identify which citation(s) a linguistic expression is attributable to. For tasks of interest to us, it is not enough to only consider anaphoric references to entire papers; authors often make statements comparing/using/criticising *aspects* or *subparts* of cited work. We therefore consider a far wider range of markables than Kim and Webber's single pronoun *"they"*.

Our attribution task differs from the traditional anaphora resolution task in that we have a fixed list of possible referents (the reference list items, Current-Paper or No-Specific-Paper), rather than a much larger set of potential references. Also, we do not form co-reference chains; we attribute a referring expression directly to one or more referents. Ours is therefore a multi-label classification task, where the citations, Current-Paper and No-Specific-Paper are the labels, and where one or more labels are assigned to each markable.

We evaluate intrinsically by comparing to human-annotated attribution, and extrinsically by showing that automatically acquired knowledge about scientific attribution improves performance on a discourse classification task—*Argumentative Zoning* (Teufel and Moens, 2002), where sentences are labelled as one of {Own, Other, Background, Textual, Aim, Basis, Contrast} according to their role in the author's argument.

We describe our data in §3 and methodology in §4, discuss evaluation metrics in §5, and evaluate intrinsically in §6 and extrinsically in §7.

## 3 Data

We used data from the CmpLg (Computation and Language archive; 320 conference articles in computational linguistics). The articles are in XML format.

We produced an annotated corpus (10 articles, 4290 data points, i.e., markables) based on written guidelines. The task was found to be quite intuitive by our annotators, and this was reflected in high agreement - Krippendorff's alpha[2] of more than 0.8 (2 annotators, 3 papers, 1429 data points) on the attribution task. The distribution of classes was, as expected, quite skewed: 69% of markables are attributable to

[1] We use a list of around 40 research methodology related nouns from Teufel and Moens (2002), such as e.g., *"study, account, investigation, result"* etc. These are nouns we are particularly interested in.

[2] see description in §5.2

CURRENT-PAPER, 7% to no specific paper and 24% to specific references (on average, 1.7 per reference). Details about the annotation process and human agreement figures can be found in Siddharthan and Teufel (2007).

## 4 Machine Learning Approach

We frame the attribution problem as a classification task: Given a markable (the definite description/pronoun/work noun under consideration), a binary yes/no decision is made for each cited paper, and a binary yes/no decision is made for whether the markable is attributable to the current paper. The list of labels for the markable is compiled by including all the citations for which the machine learner returns yes, and CURRENT-PAPER if the learner returns yes. If the list is empty (learner returns no for everything), the label is NO-SPECIFIC-PAPER.

Since the model for whether a markable is attributable to the current work is likely to be different from the model for whether it is attributable to a citation, we trained separate models for the two problems.

### 4.1 Deciding attribution to a citation

For each data point to be classified (called NP below), we create a machine learning instance for each reference list item by automatically computing the following features from POS-tagged text:

1. Properties of data point (NP) and the closest Citation instance (CIT) of the reference list item:
   (a) Type of NP (Definite Description/Work Noun/Pronoun)
   (b) CIT is a self Citation or not
   (c) CIT is syntactic (in running text) or parenthetical
   (d) Is CIT Hobbs' prediction (searching left–right starting from current sentence and then considering previous sentences, is CIT the first citation or reference to current work found)?
2. Distance measures:
   (a) Dist. between NP and CIT measured in words
   (b) Dist. between NP and CIT measured in sentences
   (c) Dist. between NP and CIT measured in paragraphs
   (d) Is CIT after NP in the discourse (cataphor)?
   (e) Distance between CIT and the closest first person pronoun or *"this paper"* in words
3. Contextual:
   (a) Rank of CIT (how many other reference list items are closer)
   (b) Number of times CIT is cited in the paragraph
   (c) Number of times CIT is cited in the whole paper
   (d) Current Section heading (this feature has 5 values: Introduction, Methods, Results, Conclusions, Unrecognised)
4. Agreement:
   (a) Agreement Number (He/She & single author non-self citation)
   (b) Agreement Person (First & Current/Self Citation, Third and Not-Current)

We have a chicken and egg problem with calculating the distance of a reference to current work in 2(e). Unlike citations, these are not unambiguously marked in the text. We calculate distance from the closest first person pronoun (even though these could possibly refer to a self citation, rather than current work) or the phrase "this paper", which can again refer to other citations but predominantly refers to current work.

### 4.2 Deciding attribution to current work

We use the same features for the second classifier that makes the decision on whether the data point refers to CURRENT-PAPER, with the following changes: Features 1(b,c) are removed as they are meaningless; 1(d) checks Hobbs' prediction for a first person pronoun/"this paper", rather than CIT; in 2(a–d), the distance is measured between the closest first person pronoun/"this paper" and the markable, rather than a citation and the markable; similarly, in 3(b,c) we count instances of first person pronoun/"this paper"; for 2(e), we now calculate the distance of the closest citation instance. In short, the same features are used, but current work and citations are swapped.

## 5 Evaluation Metrics

We consider two evaluation metrics. The first is the scoring system used for the co-reference task in the Message Understanding Conferences MUC-6 and MUC-7. The second is Krippendorff's $\alpha$. We briefly discuss both below.

### 5.1 The MUC-6/MUC-7 Metric

The MUC-6/MUC-7 Co-reference evaluation metric (Vilain et al., 1995) works by comparing co-reference classes across two annotated files. Calling one annotation the "model" and

the other the "system", for each co-reference class $S$ in the model, $c(S)$ is the minimal number of co-reference links needed to generate the class (this is one less than the cardinality of the class; $c(S) = |S| - 1$). $m(S)$ is the number of "missing" links in the system annotation relative to the co-reference class as marked up in the model. In other words, this is the minimum number of co-reference links that need to be added to the system annotation to fully generate the co-reference class $S$ in the model. Recall error is then $RE(S) = m(S)/c(S)$ and Recall is $R(S) = 1 - RE = \frac{c(S)-m(S)}{c(S)}$. Recall for the entire file (or set of files) is calculated by summing over all co-reference classes in the model:

$$R = \frac{\sum_i c(S_i) - m(S_i)}{\sum_i c(S_i)}$$

Precision ($P$) is calculated by swapping the model and system and the f-measure ($F = 2R \times P/(R + P)$) is symmetric with respect to both annotations.

### 5.2 Krippendorff's Alpha

We follow Passonneau (2004) and Poesio and Artstein (2005) in using Krippendorff (1980)'s $\alpha$ metric to compute agreement between annotations. The advantage of $\alpha$ over the more commonly used $\kappa$ metric is that $\alpha$ allows for partial agreement when annotators assign multiple labels to the same markable; in this case calculating agreement on a markable requires a more graded agreement calculation than the "1 if sets are identical and 0 otherwise" provided for by $\kappa$. Krippendorff's $\alpha$ measures disagreement, and allows for the use of distance metrics to calculate partial disagreement. Following Passonneau, we present results using four distance metrics:

1. (N)ominal: Two sets have distance $N = 0$ if they are identical and $N = 1$ if they are not. $\alpha$ calculated using the nominal distance metric is equivalent to $\kappa$.
2. (J)accard: Two sets $A$ and $B$ have distance $J = 1 - |A \cap B|/|A \cup B|$. In other words, the distance between two sets is larger, the smaller their intersection and the larger their union.
3. (D)ice: Two sets $A$ and $B$ have distance $D = 1 - 2 \times |A \cap B|/(|A| + |B|)$. In practice, the Dice distance metric behaves similarly to the Jaccard metric, but tends to be smaller, resulting in slightly higher $\alpha$.
4. (M)ASI: This is the Jaccard distance $J$ weighted by a monotonicity distance $m$ where, $m = 0$ if two sets are identical; $m = 0.33$ if one is a subset of the other; $m = 0.67$ if the intersection and the two set differences are all non-null; $m = 1$ if the two sets are disjoint. Formally, the MASI metric is $M = m \times J$.

As an example, consider two sets $\{a, b, c\}$ and $\{b, c, d\}$. The distances between these sets are $N = 1$, $J = 1-2/4 = 0.5$, $D = 1-2\times2/(3+3) = 0.33$ and $M = 0.67 \times 0.5 = 0.33$.

Krippendorff's $\alpha$ is defined as $\alpha = 1 - D_o/D_e$, where $D_o$ is the observed disagreement and $D_e$ is the disagreement that is expected by chance:

$$D_o = \frac{1}{c(c-1)} \sum_j \sum_k \sum_{k'} n_{jk} n_{jk'} d_{kk'}$$

$$D_e = \frac{1}{c(c-1)} \sum_k \sum_{k'} n_k n_{k'} d_{kk'}$$

In the above formulae, $c$ is the number of coders, $n_{jk}$ is the number of times item $j$ is classed as category $k$, $n_k$ is the number of times any item is classed as category $k$ and $d_{kk'}$ is the distance between categories $k$ and $k'$.

Like $\kappa$, Krippendorff's $\alpha$ is 1 when there is perfect agreement, 0 when the observed agreement is only what was expected by chance, negative when observed agreement is less than expected by chance and positive when observed agreement is greater than expected by chance.

## 6 Intrinsic Evaluation Results

We ran a machine learning experiment using 10-fold cross-validation and the memory-based learner IBk[3] (with k=6), using the WEKA toolkit (Witten and Frank, 2000). The performance is shown in Tables 1 and 2. To position these results we compare them with three baseline lower bounds and the human performance upper bound in Table 3. We use three baselines:

[3]Memory based learning gave better results on this task than other learners (NB, HNB, IBk, J48, cf. § 7.3.

| Paper | Items | $\alpha$-N | $\alpha$-J | $\alpha$-D | $\alpha$-M | %A* |
|---|---|---|---|---|---|---|
| 0003055 | 446 | .601 | .606 | .607 | .610 | 85% |
| 0005006 | 446 | .670 | .704 | .711 | .715 | 81% |
| 0005015 | 462 | .679 | .696 | .701 | .706 | 81% |
| 0005025 | 277 | .707 | .707 | .707 | .707 | 86% |
| 0006011 | 393 | .766 | .771 | .772 | .775 | 88% |
| 0006038 | 578 | .551 | .568 | .573 | .578 | 79% |
| 0007035 | 393 | .570 | .590 | .600 | .609 | 90% |
| 0008026 | 449 | .700 | .700 | .700 | .700 | 87% |
| 0001001 | 420 | .564 | .565 | .569 | .571 | 88% |
| 0001020 | 429 | .730 | .778 | .790 | .801 | 88% |
| AVG. | 429 | .654 | .669 | .673 | .677 | 85% |

*% Agreement, the conservative estimate measured using the Nominal metric

Table 1: Agreement with Human Gold Standard

- BASE$_M$ (Major Class): All data points are labelled `CURRENT-WORK`
- BASE$_P$ (Previous): Data points are tagged with the most recent label
- BASE$_H$ (Hobbs' Prediction): Data points are tagged with the label found by Hobbs' (1986) search (Search left to right in each sentence, starting from current sentence, then considering previous sentences)

As Table 3 shows, our machine learning approach performs much better than the baselines on all the agreement metrics, and is indeed closer to human performance than to any of the baselines. The MUC evaluation appears to produce highly inflated results on our task – when there is a small set of co-reference classes and one of these classes contains 70% of data points, it takes only a small number of missing links to correct annotations. This results in unreasonably high values, particularly for the majority class baseline of labelling every data point as CURRENT-PAPER. We believe that the $\alpha$ metrics provide a much more realistic estimate of the difficulty of the task and the relative performances of different approaches.

Table 4 shows the performance of the machine learner for each of the three types of linguistic expressions considered. Pronouns are the easiest to resolve, with on average 90% resolved correctly (an agreement with the human gold standard of $\alpha = .71$). This drops to 85% ($\alpha = .68$) for definite descriptions and demonstratives, and further to 78% ($\alpha = .63$) for remaining work nouns (i.e., those not already in a definite noun phrase).

| Paper | No. Classes | Recall | Precision | F |
|---|---|---|---|---|
| 0003055 | 14 | .934 | .886 | .910 |
| 0005006 | 17 | .875 | .870 | .872 |
| 0005015 | 19 | .897 | .876 | .886 |
| 0005025 | 16 | .903 | .874 | .888 |
| 0006011 | 14 | .942 | .909 | .925 |
| 0006038 | 25 | .905 | .893 | .899 |
| 0007035 | 18 | .957 | .926 | .941 |
| 0008026 | 9 | .966 | .962 | .964 |
| 0001001 | 14 | .949 | .908 | .928 |
| 0001020 | 18 | .924 | .926 | .925 |
| TOTAL | 164 | .924 | .903 | .913 |

Table 2: Evaluation using MUC-6/7 software

| Algo | $\alpha$-N | $\alpha$-J | $\alpha$-D | $\alpha$-M | %Agr* | muc-f |
|---|---|---|---|---|---|---|
| Base$_M$ | .002 | .001 | .001 | .001 | 69% | .934 |
| Base$_P$ | -.101 | -.083 | -.081 | -.077 | 19% | .894 |
| Base$_H$ | .387 | .397 | .399 | .407 | 72% | .910 |
| **IBk** | **.654** | **.669** | **.673** | **.677** | **85%** | **.913** |
| Hum** | .806 | .808 | .808 | .809 | 91% | .965 |

*% Agreement, the conservative estimate measured using the Nominal metric

**Agreement between two human annotators over a subset of the corpus (3 files, 1429 data points)

Table 3: Comparison with Baselines and Human Performance (Averaged results)

While all the features contributed to the reported results, the most important features (in terms of information gain) for deciding attribution to a citation were the paragraph level citation count 3(b), the distance features 2(a,b,c,d), the rank 3(a) and the Hobbs' prediction 1(d). The most important features for deciding attribution to the current paper were the distance features 2(a,c,e), the rank 3(a) and the Hobbs' prediction 1(d).

## 7 Extrinsic Evaluation

To demonstrate the use of automatic scientific attribution classification, we studied its utility for one well known discourse annotation task: Argumentative Zoning (Teufel and Moens, 2002). Argumentative Zoning (AZ) is the task of applying one of seven discourse level tags (Figure 1) to each sentence in a scientific paper.

These categories model several aspects of scientific papers: from the distinction of segments by who an idea is attributed to (`Own` – `Other` – `Background`), to the judgement of how the au-

| Paper | Pronouns | | Definites | | Work Nouns | |
|---|---|---|---|---|---|---|
| | $\alpha_M$ | $\%_N$ | $\alpha_M$ | $\%_N$ | $\alpha_M$ | $\%_N$ |
| 0003055 | .746 | 94% | .556 | 83% | .735 | 87% |
| 0005006 | .846 | 91% | .703 | 85% | .700 | 78% |
| 0005015 | .662 | 83% | .692 | 79% | .787 | 86% |
| 0005025 | .804 | 89% | .717 | 87% | .514 | 78% |
| 0006011 | .824 | 91% | .807 | 91% | .615 | 76% |
| 0006038 | .603 | 90% | .609 | 81% | .430 | 66% |
| 0007035 | .577 | 94% | .507 | 91% | .770 | 87% |
| 0008026 | .678 | 88% | .726 | 87% | .551 | 78% |
| 0011001 | .562 | 97% | .633 | 87% | .377 | 81% |
| 0011020 | .792 | 90% | .798 | 92% | .808 | 89% |
| AVG. | .709 | 90% | .675 | 85% | .629 | 78% |

Table 4: Results for different markable types

| Category | Description |
|---|---|
| `Background` | Generally accepted background knowledge |
| `Other` | Specific other work |
| `Own` | Own work: method, results, future work |
| `Aim` | Specific research goal |
| `Textual` | Textual section structure |
| `Contrast` | Contrast, comparison, weakness of other solution |
| `Basis` | Other work provides basis for own work |

Figure 1: AZ Annotation scheme

thors relate to other work (`Contrast` – `Basis`) to the rhetorical status of high-level discourse goals (statement of `Aim`; overview of section structure (`Textual`)). Some of these categories (`Background`, `Other` and `Own`) occur in zones that span many sentences. Other categories typically occur in short zones, often just a single sentence (`Textual`, `Aim`, `Contrast`, `Basis`).

In all work to date, classification of sentences into one of the AZ categories has been performed on the basis of features extracted from within the sentence, and a few contextual features such as section heading and location in document. Scientific attribution links previously unresolved noun phrases or pronouns in the sentence to citations. As this provides the machine learner with more information, AZ results should improve.

### 7.1 AZ Data

The evaluation corpus used is the one from Teufel and Moens (2002). It consists of 80 conference papers in computational linguistics, containing around 12000 sentences. Each of these is manually tagged as one of {OWN, OTH, BKG, BAS, AIM, CTR, TXT}. The reliability observed is reasonable (Kappa=0.71)).

### 7.2 Features

Following in Teufel and Moens (2002), we used supervised ML using features extracted by shallow processing (POS tagging and pattern matching):

- **Lexical (cue phrase) features** consist of three features: the first models occurrence of about 1700 manually identified scientific cue phrases (such as "in this paper"). The cue phrases are classified into semantic groups. The second models the main verb of the sentence, by lookup in a verb lexicon organised by 13 main clusters of verb types (e.g. "change verbs"), and the third models the likely subject of the sentence, by classifying them either as the authors, or other researchers, or none of the above, using an extensive lexicon of regular expressions.
- **Content word features** model occurrence and density of content words in the sentences, where content words are either defined as non-stoplist words in the subsection heading preceding the sentence, or as words with a high TF*IDF score.
- **Linguistic features** include (complex) tense, voice, and presence of an auxiliary.
- **Citation features** detect properties of formal citations in text, such as the occurrence of authors' names in text, the position of a citation in text, and whether the citation is a self citation (i.e. includes any of the authors of the paper itself).
- **Location features:** Rhetorical roles are expected at certain places in the document, for instance, background sentences are more likely to occur at the beginning of the text, and goal statements often occur after about a fifth of the paper. We model this by splitting the text into ten segments and assigning each sentence to the segment it is located in. We also use the section heading as a contextual feature.

Some categories tend to occur in blocks (e.g., OWN, OTHER, BACKGROUND), and the context in terms of the label of the previous sentence has good predictive value. We model this (the

| Learner | kappa | | Macro-F | |
|---|---|---|---|---|
| | No Attrib | With Attrib | No | With |
| NB | .45 | .46 | .53 | .53 |
| HNB | .42 | .45 | .51 | .53 |
| IBk | .34 | .36 | .39 | .39 |
| J48 | .38 | .41 | .41 | .48 |
| Stacking | .45 | .48 | .51 | .53 |

Table 5: Improvement on AZ from using automatic scientific attribution classification.

so-called **History feature**) by running the classifier twice, and including the prediction for the previous sentence as a feature the second time.

Due to practical considerations, we obtained our linguistic features using the RASP part of speech tagger (Briscoe and Carroll, 1995), when in previous work we used the LT TTT (Grover et al., 2000). We would not expect this to influence results much, however. Another difference is that we use around 1700 additional cue phrases acquired from previous work on another discourse task[4] (Teufel et al., 2006).

In addition to these features, we use four features obtained from the scientific attribution task described in this paper:

**Scientific Attribution Features**:

- Whether there is any reference to current work in the sentence
- Whether there is any reference to any specific citation in the sentence
- Whether there is any reference in the sentence to work that is in neither the current paper nor any specific citation
- Which of these, if any, is in subject position

Our aim is to explore whether these features obtained from the scientific attribution task influence machine learning performance on AZ.

### 7.3 AZ results

We ran five different machine learners with and without the four scientific attribution features (c.f., §7.2). Note that our labelled data for the attribution task does not overlap with the 80 papers in the AZ corpus, and all attribution predictions used in features for this AZ experiment are obtained entirely from unseen (and indeed

[4]These cues are acquired manually from files that are not part of the AZ evaluation corpus.

| | Without Attribution Features | | | | | | |
|---|---|---|---|---|---|---|---|
| | Aim | Ctr | Txt | Own | Bkg | Bas | Oth |
| P | .44 | .42 | .52 | .84 | .46 | .34 | .47 |
| R | .61 | .30 | .68 | .88 | .45 | .37 | .37 |
| F | .52 | .35 | .59 | .86 | .46 | .35 | .42 |
| Correctly Classified Instances | | | | | 73.0% | | |
| Kappa statistic | | | | | 0.45 | | |
| Macro-F | | | | | 0.51 | | |

| | With Attribution Features | | | | | | |
|---|---|---|---|---|---|---|---|
| | Aim | Ctr | Txt | Own | Bkg | Bas | Oth |
| P | .57 | .42 | .57 | .84 | .44 | .40 | .55 |
| R | .61 | .27 | .66 | .90 | .47 | .43 | .42 |
| F | .59 | .33 | .61 | .87 | .46 | .41 | .47 |
| Correctly Classified Instances | | | | | 74.7% | | |
| Kappa statistic | | | | | 0.48 | | |
| Macro-F | | | | | 0.53 | | |

Table 6: Best AZ results using Stacked classifier: with and without Attribution Features.

unlabelled) data based on the model learnt on 10 papers (c.f., §6). The learners we used (with default WEKA settings) are:

- NB: Naive Bayes learner
- HNB: Hidden Naive Bayes learner
- IBk: Memory based learner
- J48: Decision tree based learner
- STACKING: combining NB and J48 classifiers with the stacking method

As mentioned under *History features* above, we run each learner twice, the second time including the machine learning prediction for the previous sentence (note that in Teufel and Moens (2002) we noticed a slight improvement in performance for NB when doing so (between .005 and .01 on both Kappa and Macro-F)). We found an improvement from including the four reference features with all the learners, as shown in Table 5.

| | Aim | Ctr | Txt | Own | Bkg | Bas | Oth |
|---|---|---|---|---|---|---|---|
| P | .44 | .34 | .57 | .84 | .40 | .37 | .52 |
| R | .65 | .20 | .66 | .88 | .50 | .40 | .39 |
| F | .52 | .26 | .61 | .86 | .44 | .38 | .44 |
| Correctly Classified Instances | | | | | 72.5% | | |
| Kappa statistic | | | | | 0.45 | | |
| Macro-F | | | | | 0.50 | | |

Table 7: Teufel and Moens (2002)'s best AZ results (Naive Bayes Classifier).

For a more detailed view of where the improvement comes from, refer to Table 6, that shows precision, recall and f-measure per category for our best learner. The biggest improvements from using attribution features are for the categories OTHER, AIM and BAS. The improvement in OTHER was to be be expected, as this zone is directly related to the attribution classification. The large improvements in AIM and BAS is good news, as these are amongst our most informative rhetorical categories for downstream tasks. Our best results of Kappa=0.48 and Macro-F=0.53 are better than the best previously published results on task (Kappa=0.45 and Macro-F=0.50 in Teufel and Moens (2002)). Our results improve upon the results of Teufel and Moens (2002) (reproduced in Table 7) – both overall and for each individual category.

## 8 Conclusions

We have described a new reference task - deciding scientific attribution, and demonstrated high human agreement ($\alpha > 0.8$) on this task. Our machine learning solution using shallow features achieves an agreement of $\alpha_M = 0.68$ with the human gold standard, increasing to $\alpha_M = 0.71$ if only pronouns need to be resolved. We have also demonstrated that information about scientific attribution improves results for a discourse classification task (Argumentative Zoning).

We believe that similar improvements can be achieved on other discourse annotation tasks in the scientific literature domain. In particular, we plan to investigate the use of scientific attribution information for the citation function classification task.

### Acknowledgements

This work was funded by the EPSRC project SciBorg (EP/C010035/1, Extracting the Science from Scientific Publications).

## References

S. Bradshaw. 2003. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proc. of ECDL*.

T. Briscoe and J. Carroll. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proc. of IWPT-95*, Prague / Karlovy Vary, Czech Republic.

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

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

C. Grover, C. Matheson, A. Mikheev, and M. Moens. 2000. LT TTT - A flexible tokenisation tool. In *Proc. of LREC-00*, Athens, Greece.

W. Hersh, R. Bhuptiraju, L. Ross, P. Johnson, A. Cohen, and D. Kraemer. 2004. Trec 2004 genomics track overview. In *Proc. of TREC*.

J. Hobbs. 1986. Resolving Pronoun References. In *Readings in Natural Language*, Grosz, B., Sparck-Jones, K. and Webber, B. (eds.) Morgan Kaufman.

Y. Kim and B. Webber. 2006. Automatic reference resolution in astronomy articles. In *Proc. of 20th International CODATA Conference*, Beijing, China.

K. Krippendorff. 1980. *Content Analysis: An introduction to its methodology*. Sage Publications, Beverly Hills.

H. Nanba and M. Okumura. 1999. Towards multi-paper summarization using reference information. In *Proc. of IJCAI-99*.

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

R. Passonneau. 2004. Computing reliability for coreference annotation. In *Proc. of LREC-04*, Lisbon, Portugal.

M. Poesio and R. Artstein. 2005. Annotating (anaphoric) ambiguity. In *Proc. of the Corpus Linguistics Conference*, Birmingham, UK.

A. Siddharthan and S. Teufel. 2007. Whose idea was this? Deciding attribution in scientific literature. In *Proc. of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC'07)*, Lagos, Portugal.

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

S. Teufel, A. Siddharthan, and D. Tidhar. 2006. Automatic classification of citation function. In *Proc. of EMNLP-06*, Sydney, Australia.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the 6th Message Understanding Conference*, San Francisco.

I. Witten and E. Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.