Implicit Reference to Citations: A study of astronomy papers

Yunhyong Kim
Digital Curation Centre (DCC)
&
Humanities Adavanced Technology Information Institute (HATII)
University of Glasgow

Bonnie Webber
School of Informatics
University of Edinburgh

October 5, 2006

Abstract

The research in this paper presents results in the automatic classification of pronouns within articles into those which refer to cited research and those which do not. It also discusses the automatic linking of pronouns which do refer to citations to their corresponding citations. The current study focussed on the pronoun *they* as used in papers in Astronomy journals. The paper describes a classifier trained on maximum entropy principles using features defined by the distance to preceding citations and the category of verbs associated to the pronoun under consideration.

1 Introduction

Scientific information comes in many different forms — databases, scientific articles, memos, notes or emails sent to colleagues on the spur of the moment. To understand scientific text, it is important to put it into context by linking it to related information. This leads us to contemplate methods for drawing out links between scientific texts, as in citation linking, i.e. linking an article to the other articles that it cites. In academic research, the citation index of an article, defined using the number of times the article is cited by other articles, plays a role in determining the impact of the research in the paper. In fact, given no other information, to acquire fast-track knowledge of core work in a chosen field, it often pays to look at articles which are cited frequently. To link articles by way of citations can therefore be quite useful. In fact, publishers of on-line journals (e.g. [3], [17]) have already implemented active hyper-links to articles cited within the articles that appear in their journal, and services, such as Google Scholar and CiteSeer (e.g. [4]), list other papers which cite and are cited by the retrieved article. Discussions of techniques in citation linking can be found in papers such as [2] and [9].

The work in the present paper is motivated by a conviction that recognising implicit references to citations, as well as recognizing explicit citations, can also lead to a better understanding of the document. It has already been suggested that, to capture what is significant about a paper, one can examine the sentences in which citations to that paper appear [14]. It follows that it may also be quite useful to identify and link implicit references to citations (e.g. via pronouns or anaphoric NPs) to mine the corresponding sentences for even deeper context; it is often within the sentences of subsequent references that the content of the cited articles is introduced. Linking such references to their corresponding citation would also help in (1) measuring the salience of articles (based on the intuition that articles referenced many times are more likely to be important and closely related to the current article than those which are not), (2) summarising articles [18], and (3) facilitating information mining because keywords of the topic often appear within reference sentences [14].

The present paper attempts a solution to recognizing and linking implicit references by viewing it as an anaphora resolution problem. An anaphor A of an antecedent B is a phrase or word whose interpretation depends on B. For instance, in the text

"John was late for school. He had taken the wrong bus.",

He is an anaphor of the antecedent *John*. Linguistically speaking, the references to citations in a research article can be thought of as anaphors whose antecedents are citations. In this context, anaphor resolution, i.e., detecting the anaphors and resolving them to the correct citation, is equivalent to linking implicit references to their corresponding citations.

In this work, we have focussed on implicit citations using the personal pronoun *they*. In astronomy, most published research seems to be collaborative, due to the fact that observational astronomy involves a range of skills in instruments, observation, computer programs and data analysis. Implicit citations to a paper are thus commonly in terms of implicit reference to its multiple authors. (In a corpus of 35 articles, we found only seven instances of the pronoun *he* anaphoric to citations, compared with sixty instances of the pronoun *they*.) While the singular pronoun *it* can be used to refer to the paper itself, we did not consider *it* in this initial study.

There are other studies on resolving anaphors (e.g. [11], [12], [8]), on distinguishing between referential and non-referential *it* (e.g. [4], [13]), and on coreference resolution (e.g. [15]). The current paper, however, focusses on citations and references to citations. We did not concern ourselves with resolving instances of referential *they* after deciding that they were not anaphoric to citations.

The corpus used in this research came from articles in Monthly Notices of the Royal Astronomical Society (MNRAS). The articles were retrieved in HTML format with hyperlinks to citations linked by Blackwell Publishing [3] already in place.[1]

Although the work relates to scientific articles one can also envision leaving the arena of scientific articles to a wider domain of texts to apply the detection method outlined in this paper to mine noun phrases which constitute informal references to scientific work. This may provide a means by which to link scientific discussions in informal contexts such as emails or letters to formal scientific work, leading to a contextually richer information network.

[1]Some citations were not properly linked and were not corrected before carrying out the experiments in this paper. The performance of our resolution algorithm, which depended on the citations already being linked, would improve with better citation linking.

2 Scope of references

Pronominal references to citations can appear in several forms, including personal pronouns (*I*, *he*, *she*, *it*, *we* and *they*, and their corresponding inflected forms), definite noun phrases (e.g. "the paper", "the study") and names or other incomplete citations.

As mentioned in Section 1, we have concentrated here on the personal pronoun *they*, as most work in astronomy is collaborative, making *they* the most popular form of pronominal reference to citations. *They* can also appear inflected as *them* or *their*, although *them* occured with only 24% the frequency of *they* or *their* in the corpus, and only one instance was identified as referring to a citation. Thus, it was not included in the study. While *their* and *they* occurred with almost equal frequency (141 vs. 142 instances), *their* functioned less frequently as a citation reference (18% vs. 30%). When *their* did refer to a citation, 65% of the time, it was to an explicit citation in the same sentence. Since our initial interest is in identifying the full set of sentences that contain some citation (either explicitly or implicitly), only 6% of the instances of *their* would bring another sentence into this set. While this is not insignificant, we again put off further analysis of *their* to future work.

The pronoun *they* has multiple uses in astronomy. It can refer to specific elements in the domain (Example 1) or to a citation's authors (Example 2):

1. These correlated maps satisfy a desirable property since *they* are model independent.

2. *They* showed that the most probable hypertorus was 1.2 times larger than the horizon scale.

They can also be used as in Example 3 to make a general statement which does not refer to any specific research.

3. In experiments of the type described here, *they* do not include such data.

More generally, *they* can even refer back to a singular entity to avoid gender specification[2]. These multiple uses make the task of identifying instances of *they* which refer to citations difficult but interesting.

The data used in the development and testing of the system was extracted from research articles available in Hyper Text Markup Language (HTML) at the Monthly Notices of the Royal Astronomical Society (MNRAS). For text processing articles with pre-linked citations, HTML seemed like a reasonable choice. Although there are articles in PDF with links to citations in place, this is less common, and extracting text from PDF files is technically more complicated and often prohibited due to security issues. The selection of MNRAS as the source was mainly due to the fact that other astronomical journals were not publicly accessible without a subscription. There are two training datasets two test data sets:

- Training data:
  - **main training data set** TrA consisting of 377 samples of *they* retrieved from 51 articles of the MNRAS.
  - **supplemental training data set** TrSup consisting of 64 samples of *they* from 9 articles of the MNRAS
- Test data:

[2]Someone left the door open last night. *They* had better be careful in the future.

- **developmental test data set** TeA-1 consisting of 205 samples extracted from 35 articles of the MNRAS
- **fresh test data set** TeA-2 consisting of 211 samples of *they* from 35 articles.

There were 81 positive samples (i.e. eighty one *they*s which referred to a citation) in the 377 samples of *they* collected from TrA. For each dataset, twenty to thirty percent of all *they*s were positive examples.

3 Tools

Two main tools were employed for extracting features and training the system: Maximum Entropy toolkit (MaxEnt)(cite here) and the CANDC part-of-speech tagger.

MaxEnt

The statistical modeling tool used in the research of this paper was the Maximum Entropy toolkit (MaxEnt) developed by Zhang Le (cite here). As its name suggests, this toolkit is designed to train a classifier based on the maximum entropy principle. The MaxEnt toolkit was selected for the reason that the maximum entropy model makes the least number of assumptions on the independence of selected features. It has been used successfully in other language processing tasks. Maximum entropy however has been known to have symptoms such as overfitting the training data, so it will be important to supplement the current work with trials of other statistical models.

CANDC

We used the the CANDC tagger developed by Curran and Clark ([7]) to get the part-of-speech tags for the text in the astronomy articles. The tagger is trained on the Penn Treebank text using a maximum entropy model with parameters estimated by the GIS algorithm. The tagger takes tokenised text with all punctuation separated from the words by white space.The tags used to label the part-of-speech is the penn treebank set ([16]).

4 Resolving implicit references

Resolving the implicit references which refer to citations involves two steps: making a decision whether a given string refers to a citation (**Detection**), and linking strings determined to be anaphoric to the correct citation (**Resolution**).

4.1 Detection

Anaphors were identified as implicit citations with a binary classifier trained using MaxEnt on the features described below.

Features

a. Distance to the nearest preceding active citation

The inclusion of this feature is motivated by the intuition that most secondary references to citations lie close to the initial citation. The term "active citation" is defined to be a

citation which is not etirely placed within a paranthesis. The notion stems from an effort to make a distinction between citations that form part of a sentence as a syntactic unit and those that do not. For example,

- the following citations are active:
  - Smith et al. (1998) studied the distribution of ...
  - In [2], it was proved that the angular velocity of ...
- and the following citations are inactive:
  - Recent work shows that the distribution of galaxies is uniform. (Smith et al. 1998)
  - Galileo presented observational data that the earth is round ([4]).

The use of *active citations* is reflective of the fact that most citations which are discussed extensively will have a prominant syntactic role in the language. This may not be true in all domains.

The distances which we will call the nearest citation index (NCI) from the reference under consideration to the preceding citations are measured by number of sentences. If the nearest preceding active citation is in the same sentence as the reference then the distance is zero. In the experiments of this paper we cosidered the NCI on only three levels: C (close), N (neutral), and F (far). The label C was given to references for which $NCI \leq 1$, the label N to those for which $NCI \leq 5$ and F otherwise.

b. Verb categories

As noted earlier, this study focussed on sentences containing the pronoun *they*. This uninflected form occurs only in subject position in a sentence, generally followed by a verb group of the form

$$they \quad V_4 \quad V_3 \quad V_2 \quad V_1 \quad V_0 \quad \text{(possibly an object)}.$$

where V_i $(i = 0, \ldots, 4)$ have the following syntactic categories:

V_4	modal
V_3	auxillary have
V_2	auxillary be
V_1	progressive be
V_0	main verb.

For example, all verbs V_i are present in the sentence

> They may have been being calibrated.

We were interested in the presence or absence of each V_i and the form it takes, i.e.,

- the part-of-speech tag of the main verb V_0 (e.g. base form, past tense, gerund, past participle, first or third person singular present);
- the aspect of the sentence (e.g. simple, perfect, or progressive), determinable from V_3, V_1 and the inflection of V_0;

- whether the sentence is in a passive or active voice (usually distinguished by whether it takes the form *be + past participle* or not);
- the identify of each verb;
- the presence of a modal (e.g. *can*, *may*, *shall*, *will* and *do*, and related inflected forms).

For transitive verbs, we were interested in the syntactic type of its object (e.g. clausal complement or noun phrase).

The part-of-speech tag (POS tag) of the main verb can serve as partial evidence for a temporal index establishing a relationship between the time of the event described and the time of reference to the event. Events in cited papers occur before their reference time, and sentences containing such citations are expected to exhibit this tendency. In addition, the POS tag of the main verb, in combination with the presence of the auxilary verb *be*, approximates a detector of the passive voice. The intuition for including this information is that the paper or the researchers of a paper would play an active role and would not be likely to occur in the passive context. Lexical information about the verbs is meant to identify verbs used when reporting research results such as *find, argue or conclude*. The use of a frequently-used-words list or WordNet to enhance this feature was contemplated but not employed for the experiments in this paper. Also, the examples

1. *They* discovered that galaxies are uniformly distributed.(**anaphoric to citation**)
2. *They* gradually disintegrate. (**not anaphoric to citation**)
3. *They* slow down the velocity. (**not anaphoric to citation**)

suggest that for transitive verbs, the type of object may also be evidence for the referent of the pronoun *they*.

4.2 Resolution

Once a string has been identified as a reference to a citation, we must link the reference to the corresponding citation by locating the correct citation. We have used a simple rule-based algorithm which uses the notion of an *active citation* defined at the begining of Section 4.1: A citation is said to be *active* if it is not entirely within parentheses. We use the term *inactive citation* to refer to citations which are not active.

The resolution algorithm can be described as follows:

- if there is an active citation in the same paragraph, link to the nearest preceding active citation
- if there is no active citation in the same paragraph, link to the nearest preceding inactive citation in the same paragraph
- if there are no citations in the same paragraph then move to the preceding paragraph and repeat the process until a citation is found or the beginning of the document is reached

The algorithm is based on the belief that pronominal references to citations do not lie very far away from the original citation.

5 Preprocessing: for detection

Before extracting the features necessary for the training of the detection classifier, we used the following process to format the text:

1. filter out all sentences that do not contain the word *they*
2. tokenize each sentence so that all words are separated by exactly one white space
3. replace all special symbols and punctuation marks with words describing the items replaced (e.g. a full stop will be replaced with the word "PERIOD").
4. tag each sentence with part-of-speech using *pos* from CANDC ([7])
5. make some minor corrections on the result of CANDC so that
   - obvious mathematical symbols are given the part-of-speech (POS) tag 'sym'
   - the string "fig (abbreviation of figure)" is given POS tag 'nn'
   - the string "spectra" is given POS tag 'nns'
   - the strings "all", "both", and "each" are given POS tag 'pdt'
6. apply a sentence simplifier which eliminates most of the sentence except
   - the position of the word *they*,
   - the positon, lexicon, and part-of-speech of verbs following the word *they*,
   - position of any existing words tagged as a preposition by CANDC,
   - the position of any noun phrase,
   - the position of clausal complement "that".

The capture of verb categories mentioned in Section 4.1 were approximated by using regular expressions to extract any verbs (and their corresponding POS tags) in the sentence, and by checking for sequences of the form

1. they + verb
2. they + verb + that
3. they + verb + verb + noun phrase
4. they + modal.

When the sentence contained less than two verbs then the label *NONE* was used in place of the lexicon. Note that sentences of form 1 are inclusive of sentences of form 2 and 3, hence, features are not independent of each other.

We have not made a distinction between all combinations resulting from the existence or absence of V_i in the basic sentence structure mentioned of Section 4.1, since it was rare for there to be more than two verbs following *they* in the sentence. Also rare were instances of clausal complements where auxilary verbs or modals were present. Given the small training data set, it seemed acceptable to disregard these rare cases for a first statistical approximation.

6 Experiments for detection

Anaphor detection was tested using three binary labellers trained using the maximum entropy toolkit MaxEnt (cite here) on the features specified below for each system.

- System A (trained on TrA):
  - the Nearest Citation Index (NCI),
  - whether or not the verb takes a clausal complement as an object
  - the lexicon of the first, second, and third verb.
  - part-of-speech tag of the main verb
- System B (trained on TrA):
  - the NCI
  - whether or not the verb takes a clausal complement as an object
  - the lexicon of the first verb
  - part-of-speech tag of the main verb
  - labels resulting from System A
- System C (trained on TrA and TrSup together):
  - whether or not the verb takes a clausal complement as an object
  - whether or not the verb takes a noun phrase as an object
  - whether there is modal present
  - whether or not there is exactly one auxillary verb and a main verb which takes a noun phrase as an object
  - the part-of speech tag of the main verb

In the experiments to follow, we will compare System A to a combined system (shown in Figure 1) that uses System A to divide the samples into those which System A labelled as anaphoric to citations (*A-yes group*) or not (*A-no group*), with the *A-yes group* then re-labelled by System B and the *A-no group* re-labelled by System C to get the final labels. The choice of features used in each system and the overall structure of the tree was a result of trial and error: various different combinations were tried and this gave the best result.

7 Results

In the evaluation of the results to follow, we use three indices that are standard in a classification tasks: accuracy, precision and recall. Let N be the total number of documents in the test data, N_c the number of documents in the test data which are in class C, T the total number of correctly labelled documents in the data independent of the class, T_c the number of true positives for class C (documents correctly labelled as class C), and F_c the number of false positives for class C (documents labelled incorrectly as class C). Accuracy is defined to be $A = \frac{T}{N}$ while precision and recall for each class C is defined to be $P_c = \frac{T_c}{(T_c+F_c)}$ and $R_c = \frac{T_c}{N_c}$ respectively.

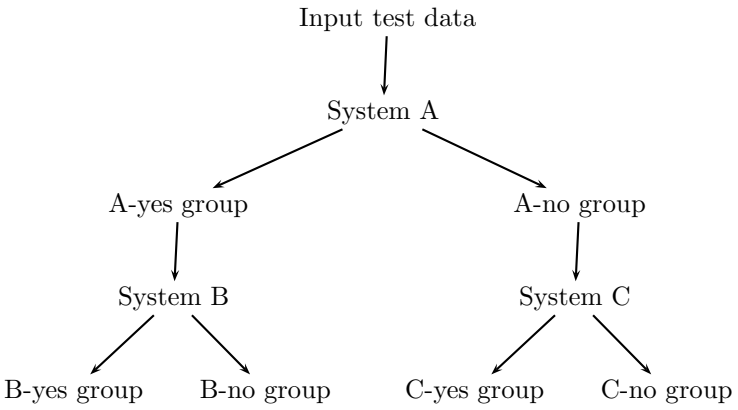

Figure 1: **Combined system decision tree**

7.1 Detection

To have a standard against which to gauge the system described in this paper, we had an Astronomer label 100 sample from the developmental test data TeA-1 and fresh test data TeA-2. The results are in Table 1. The result for the detection of anaphoric *they* is given in

System	Accuracy	Precision	Recall
Human (100 from TeA-1)	97%	92.31%	100%
Human (100 from TeA-2)	96%	not available*	100%
*two were labelled ambiguous			

Table 1: On the developmental data (top) and fresh test data (bottom): human annotator

Table 2. The Table displays the accuracy, precision and recall of *System A* as a standalone system and the *Combined System* of *A,B and C* (Figure 1). The results show that, although *System A* has very good precision, recall is not as good as the *Combined System* on both the developmental data (TeA-1) and fresh test data (TeA-2).

System	Accuracy	Precision	Recall
System A	94.63%	100%	81.66%
Combined	96.09%	91.93%	95%

System	Accuracy	Precision	Recall
System A	90.99%	97.87%	71.87%
Combined	92.41%	96.15%	78.12%

Table 2: **Tested on developmental test data TeA-1 (left) and on fresh test data TeA-2 (right)**

Error analysis

In this section we will examine the pronouns incorrectly labelled by the classifier. The tables presented in the section have been formatted so that the columns, in order, from left to right, denote the correct label of the pronoun, distance to preceding citations based on NCI, the presence or absence of a *that* clausal complement, the presence or absence of a noun phrase object, the presence or absence of modals, lexicon of the verbs, and the part-of-speech tag

of the main verb. The sample is given the label *Y* if it is anaphoric to a citation and *N* otherwise; the labels for the distance to citations are explained at the end of Section 4.1; the label 1 denotes presence and 0 denotes absence; and the part-of-speech tag follows the *Penn Treebank* schema ([16]).

The items in the developmental test data set incorrectly labelled by System A and the Combined System are given in Table 3 and Table 4, respectively. Note that items with verb such as *tried, find, argue, show, derive*, which we can imagine would be typical verbs used in describing the content of cited work, failed to be labelled correctly in System A, were correctly re-labelled by the Combined System. Table 5 and Table 6 give the

true label	citation	that	nx	md	vb+vb+nx	lexicon 1	lexicon 2	lexicon 3	tense
Y	F	0	1	0	0	tried	NONE	NONE	vbd
Y	F	0	1	0	0	take	NONE	NONE	vbp
Y	F	0	1	0	0	used	NONE	NONE	vbd
Y	F	0	1	0	0	developed	NONE	NONE	vbd
Y	F	0	1	0	0	find	NONE	NONE	vbp
Y	F	1	0	0	0	argue	NONE	NONE	vbp
Y	F	1	0	0	0	show	NONE	NONE	vbp
Y	F	0	1	0	0	found	NONE	NONE	vbd
Y	F	0	1	0	0	make	NONE	NONE	vbp
Y	F	0	1	0	0	condense	NONE	NONE	vbp
Y	C	0	0	0	0	derive	NONE	NONE	vbd
total: 11, false negatives: 11, false positives: 0									

Table 3: **Incorrectly labeled items: System A testing on developmental data TeA-1**

true label	citation	that	nx	md	vb+vb+nx	lexicon 1	lexicon 2	lexicon 3	tense
Y	F	0	1	0	0	take	NONE	NONE	vbp
Y	F	0	1	0	0	make	NONE	NONE	vbp
Y	F	0	1	0	0	condense	NONE	NONE	vbp
N	F	0	1	0	0	found	NONE	NONE	vbd
N	F	0	1	0	0	were	NONE	NONE	vbd
N	F	0	0	0	0	took	NONE	NONE	vbd
N	F	1	0	0	0	confirm	NONE	NONE	vbp
N	F	0	1	0	0	discarded	NONE	NONE	vbd
total: 8, false negatives: 3, false positives: 5									

Table 4: **Incorrectly labeled items: Combined System testing on developmental data TeA-1**

incorrectly labelled items for System A and the Combined Syetem, respectively, on the fresh test data set. It is clear that the Conbined system here also succeeded in correctly re-labelling items with verbs *proposed, consider, derive* as anaphoric to citation. But it seems that less used verbs which humans might easily expect to be citation-related (e.g. *report, expect, demonstrate*) have not been identified by the Combined System, perhaps due to the small size of the training data.

In addition to the pronoun *they*, we also experimented briefly with the pronoun *he* by

true label	citation	that	nx	md	vb+vb+nx	lexicon 1	lexicon 2	lexicon 3	tense
Y	C	0	0	0	0	rule	NONE	NONE	vbp
Y	N	0	1	0	0	have	NONE	NONE	vbp
Y	C	0	0	0	0	expect	NONE	NONE	vbp
Y	N	0	1	0	0	give	NONE	NONE	vbp
Y	C	0	1	0	0	report	NONE	NONE	vbp
Y	N	0	1	0	0	derive	NONE	NONE	vbp
Y	C	0	0	0	0	demonstrate	NONE	NONE	vb
Y	C	0	1	0	0	discuss	NONE	NONE	vbp
Y	F	0	1	0	0	identified	NONE	NONE	vbd
Y	C	0	1	0	0	subtract	NONE	NONE	vbp
Y	C	0	1	0	0	record	NONE	NONE	vb
Y	N	0	0	0	0	be	NONE	NONE	vbp
Y	C	0	1	0	0	compare	NONE	NONE	vbp
Y	N	0	0	0	0	separate	NONE	NONE	vbp
Y	F	0	0	0	0	proposed	NONE	NONE	vbd
Y	C	0	1	0	0	consider	NONE	NONE	vbp
Y	N	0	0	0	1	have	resolved	NONE	vbn
Y	C	0	0	0	1	have	produced	NONE	vbn
N	C	0	1	0	0	determine	NONE	NONE	vbp
total: 19, false negatives: 18, false positives: 1									

Table 5: **Incorrectly labelled items: System A testing on fresh data TeA-2**

true label	citation	that	nx	md	vb+vb+nx	lexicon 1	lexicon 2	lexicon 3	tense
Y	C	0	0	0	0	rule	NONE	NONE	vbp
Y	N	0	1	0	0	have	NONE	NONE	vbp
Y	C	0	0	0	0	expect	NONE	NONE	vbp
Y	N	0	1	0	0	give	NONE	NONE	vbp
Y	C	0	1	0	0	report	NONE	NONE	vbp
Y	C	0	0	0	0	demonstrate	NONE	NONE	vb
Y	C	0	1	0	0	discuss	NONE	NONE	vbp
Y	C	0	1	0	0	subtract	NONE	NONE	vbp
Y	C	0	1	0	0	record	NONE	NONE	vb
Y	N	0	0	0	0	be	NONE	NONE	vbp
Y	C	0	1	0	0	compare	NONE	NONE	vbp
Y	N	0	0	0	0	separate	NONE	NONE	vbp
Y	N	0	0	0	1	have	resolved	NONE	vbn
Y	C	0	0	0	1	have	produced	NONE	vbn
N	C	0	1	0	0	determine	NONE	NONE	vbp
N	F	0	1	0	0	experienced	NONE	NONE	vbd
total: 16, false negatives: 14, false positives: 2									

Table 6: **Incorrectly labelled items: Combined System testing on fresh data TeA-2**

training the system on 28 samples of *he* extracted from 50 articles and testing on 46 examples from the developmental test articles. The System A performed at an accuracy of 89.13%

and the Combined System performed even worse at an accuracy of 80.43%. However, upon re-examination of the samples, it turned out that much of the inaccuracy was due to errors made by the CANDC POS tagger, which labelled every instance of the string *He* as a pronoun: in actuallity, most of the strings *He* (39 out of 46 samples) referred to the element Helium. Having incorrectly labelled the string as a pronoun, the POS tagger went on to incorrectly label other items as verbs and nouns misleading the system completely. Upon correcting some major POS tag errors, re-training and testing, System A performed at an accuracy of 93.48% and the Combined System at 95.65%.

7.2 Resolution

Recall the algorithm described in Section 4.2. Implementing this algorithm on the *they* in TeA-1 and TeA-2 yieled the results in Table 7. It should be pointed out that the algorithm was implemented with the assumption that the document had already been linked completely for explicit citations, i.e. when looking for preceding active citations, the code looked for linked citations. In reality, there were some citations which were not linked; if, these citations had been properly linked, the accuracy on both test data would have been over 95%.

Data	Accuracy	anaphoric items	incorrect
TeA-1	90%	60	6
TeA-2	93.75%	64	4

Table 7: **Resolution only**

7.3 Detection and resolution

The worst case and best case scenario for the complete detection and linking method have been presented in Table 8. The worst case scenario assumes that there is no intersection between incorrectly labelled items and incorrectly linked items while the best case scenario assumes that the incorrectly labelled items were inclusive in the set of incorrectly linked items. The result is not conclusive but looks very promising. A trial run of the combined system (each module automated but pipelined by hand) on a randomly selected new article, which contained 15 *they*s out of which 10 referred to a citation, returned 100% accuracy.

8 Discussion and Conclusion

The results in this paper suggest that

System	Precision	Recall
System A (TeA-1)	90%	73.5%
System A (TeA-2)	91.74%	67.31%
Combined (TeA-1)	81.42%	85.5%
Combined (TeA-1)	90.13%	73.23%

System	Precision	Recall
System A (TeA-1)	100%	81.67%
System A (TeA-2)	97.87%	71.87%
Combined (TeA-1)	90.48%	95%
Combined (TeA-1)	96.15%	78.15%

Table 8: **Detection and resolution put together: worst case (left) and best case (right)**

- Anaphora resolution techniques can achieve a fair result in the task of automatically resolving implicit references to citations
- machine learning techniques using maximum entropy modeling can model a reasonable detector of implicit references anaphoric to citations
  - on a relatively small amount of training data (cf. tasks such as part-of-speech tagging).
  - on a small number of features
  - on almost sentential information only.

Inconclusive but also suggested is that the combined system improves the recall of the detection system. By conducting a more rigourous re-implementation of the system and testing on a wider variety of article within and across subjects a better understanding of the important linguistic structures involved in identifying anaphoric implicit references should be possible.

Although other issues such as the possibility of using the lexicon (not only the position) of the head noun in the noun phrases as a feature were contemplated along with using extra-sentential information to improve detection, we did not have time to implement these features. Another interesting angle to consider would be to look at the problem from the named entity recognition point of view by considering implicit references as a subset of the entity characterised by citations. Only further experimentation would show whether such an approach would be viable.

Acknowledgements

This research was conducted as part of the MSc. for Speech and Language Processing offered at the University of Edinburgh. The research was partly carried out at the Institut des Hautes Etudes Scientifiques (IHES) where one of the authors received a fellowship. Addional work that was required to produce this paper was carried out while one of the authors was being funded at the Digital Curation Centre (DCC) and the Humanities Advanced Technology Information Institute (HATII), University of Glasgow.

Note on website citations: All citations of websites were validated on 29 May 2006.

References

[1] Berger, A., Della Pietra, S. and Della Pietra, V., 1996, "A maximum entropy approach to natural language processing", Computational Linguistics, **Vol 22, Number 1**, 39-71.

[2] Bergmark, D., 2000, "Automatic Extraction of Reference Linking Information from Online Documents", Cornell Computer Science Department, Technical Report TR 2000-1821.

[3] Blackwell Publishing, http://www.blackwellpublishing.com/

[4] Boyd, A., Gegg-Harrison, W., Byron, D., 2005, "Identifying non-referential it: A machine learning approach incorporating linguistically motivated patterns", Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing, 40-47.

[5] CiteSeer, http://citeseer.ist.psu.edu/

[6] Curran, J. R. and Clark, S., 2003, "Language independent NER using a maximum entropy tagger", Proceedings of the Seventh Conference on Natural Language Learning, 164-167.

[7] Curran, James and Clark, Stephen, 2003, "Investigating GIS and Smoothing for Maximum Entropy Taggers", Proceedings, Aunnual Meeting, European Chapter of the Assoc. of Computational Linguistics, 91-98.

[8] Ge, N., Hale,J. and Charniak, E., 1998, "A Statistical Approach to Anaphora Resolution", Proceedings of the Sixth Workshop on Very Large Corpora.

[9] Hitchcock, S and Carr, L and Harris, S and Hey, J M N and Hall, W, 1997, "Citation Linking: Improving access to online journals", Second ACM International Conference on Digital Libraries, 115-122.

[10] Malouf, R., 2002, "A Comparison of Algorithms for Maximum Entropy Parameter Estimation", Proceedings of the Sixth Conference on Natural Language Learning, 2002, 49-55.

[11] Mitkov, R., 1997, "Factors in Anaphora Resolution: they are not the only things that matter. a case study based on two different aproaches", Proceeedings of the ACL'97/EACL'97 Workshop Operational Factors in Practical Robust Anaphora Resolution.

[12] Mitkov, R., Evans, R. and Orasan, C., 2000, "A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method", Proceedings of CICLing-2000, http://clg.wlv.ac.uk/papers/papers-by-author.php?authorID=3

[13] Müller, C., 2006, "Automatic Detection of Nonreferential *it* in Spoken Multi-party Dialog", 11th Conference of the European Chapter of the Associationfor Computational Linguistics.

[14] Nakov, P., Shwartz, A., Hearst, M., 2004, "Citances: Citation Sentences for Semantic Analysis of Bioscience Text", Workshop on Search and Discovery in Bioinformatics at SIGIR'04, Sheffield, UK.

[15] Ng, V., Cardie, C., 2002, "Improving Machine Learning Approaches to Coreference Resolution", 40th Annual Meeting of the Asssociation for Computational Linguistics (ACL).

[16] Penn Treebank Tag Set, http://www.cis.upenn.edu/treebank/

[17] PubMed, http://www.ncbi.nih.gov/entrez/

[18] Teufel, Simone and Moens, Marc, 2002, "Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status", Computational LInguistics, **Vol 28, Number 4**, http://www.cl.cam.ac.uk/users/sht25/publications.html

[19] Le, Zhang, 2004, *Maximum Entropy Toolkit for Python and C++*,http://www.nlplab.cn/zhangle/maxent_toolkit.htm

[20] WordNet: An Electronic Lexical Database, MIT Press, http://wordnet.princeton.edu/