Cross-Document Event Coreference: Annotations, Experiments, and Observations

Amit Bagga
General Electric Company CRD
PO Box 8
Schenectady, NY 12309
bagga@crd.ge.com
518-387-7077

Breck Baldwin
IRCS, University of Pennsylvania
3401 Walnut Street, #400C
Philadelphia, PA 19104
breck@unagi.cis.upenn.edu
215-898-0329

1 Abstract

We have developed cross document event tracking technology that extends our earlier efforts in cross document person coreference. The software takes class of events, like "resignations" and clusters documents that mention resignations into equivalence classes. Documents belong to the same equivalence class if they mention the same "resignation" event, i.e. resignations involving the same person, time, and organization. Other events evaluated include "elections" and "espionage" events. Results range from 45-90% F-measure scores and we present a brief interannotator study for the "elections" data set.

2 Introduction

Events form the backbone of the reasons why people communicate to one another. News is interesting and important because it describes actions, changes of state and new relationships between individuals. While the communicative importance of described events is evident, the phenomenon has proved difficult to recognize and manipulate in automated ways (example: MUC information extraction efforts).

We began this research program by developing algorithms to determine whether two mentions of a name, example "John Smith", in different documents actually referred to the same individual in the world. The system that we built was quite successful at resolving cross-document entoty coreference (Bagga, 98b). We, therefore, decided to extend the system so that it could handle events as well. Our goal was to determine whether events in separate documents, example "resignations", referred to the same event in the world (is it the same person resigning from the same company at the same time). This new class of coreference has proved to be more challenging.

Below we will present our approach and results as follows: First we discuss how this research is different from Information Extraction and Topic Detection and Tracking. Then we present the core algorithm for cross document person coreference and our method of scoring the the system's output. The method for determining event reference follows with presentation and discussion of results. We finish with an interannotator agreement experiment and future work.

3 Differences between Cross Document Event Reference and IE and TDT

Before proceeding further, it should be emphasized that cross-document event reference is a distinct goal from Information Extraction (IE) and Topic Detection and Tracking (TDT).

Our approach differs from both IE and TDT in that it takes a very abstract definition of an event as a starting place, for instance the initial set of documents for resignation events consists of documents that have "resign" as a sub-string. This is even less information than information retrieval evaluations like TREC. IE takes as an event description large hand built event recognizers that are typically finite state machines. TDT starts with rather verbose descriptions of events. In addition to differences in what these technologies take as input to describe the event, the goal of the technologies differ as well.

Information Extraction focuses on mapping from free text into structured data formats like database entries. Two separate instances of an event in two documents would be mapped into the database structures without consideration whether they were the same event or not. In fact, cross-document event tracking could well help information extraction systems by identifying sets of documents that describe the same event, and giving the patterns multiple chances to find a match.

Topic Detection and Tracking seeks to classify a stream of documents into "bins" based on a description of the bins. Looking at the tasks from the TDT-2 evaluation, there are examples that are more general and tasks that are more specific than our annotation. For example, the topic "Asian bailouts by the IMF" clusters documents into the same bin irrespective of which country is being bailed out. Our approach would try to more finely individuate the documents by distinguishing between countries and times. Another TDT topic involved the Texas Cat-

John Perry, of Weston Golf Club, announced his resignation yesterday. He was the President of the Massachusetts Golf Association. During his two years in office, Perry guided the MGA into a closer relationship with the Women's Golf Association of Massachusetts.

Figure 2: Extract from doc.36

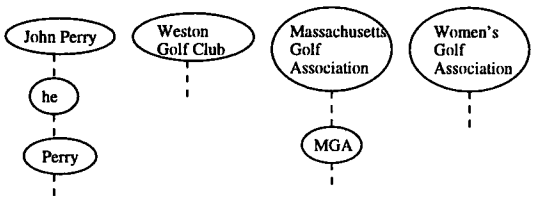

Figure 3: Coreference Chains for doc.36

tlemen's Association lawsuit against Oprah Winfrey. Given "lawsuits" as an event, we would seek to put documents mentioning that lawsuit into the same equivalent class, but would also form equivalence classes of for other lawsuits. In addition, our eventual goal is to provide generic cross-document coreference for all entities/events in a document i.e. we want to resolve cross-docuemtn coreferences for all entities and events mentioned in a document. This goal is significantly different from TDT's goal of classifying a stream of documents into "bins".

4 Cross-Document Coreference for Individuals

The primary technology that drives this research is cross-document coreference. Until recently, cross-document coreference had been thought to be a hard problem to solve (Grishman, 94). However, preliminary results in (Bagga, 98a) and (Bagga, 98b) show that high quality cross-document coreference is achievable.

Figure 1 shows the architecture of the cross-document system built. Details about each of the main steps of the cross-document coreference algorithm are given below.

- First, for each article, the within document coreference module of the University of Pennsylvania's CAMP system is run on that article. It produces coreference chains for all the entities mentioned in the article. For example, consider the two extracts in Figures 2 and 4. The coreference chains output by CAMP for the two extracts are shown in Figures 3 and 5.
- Next, for the coreference chain of interest within each article (for example, the coreference chain

Oliver "Biff" Kelly of Weymouth succeeds John Perry as president of the Massachusetts Golf Association. "We will have continued growth in the future," said Kelly, who will serve for two years. "There's been a lot of changes and there will be continued changes as we head into the year 2000."

Figure 4: Extract from doc.38

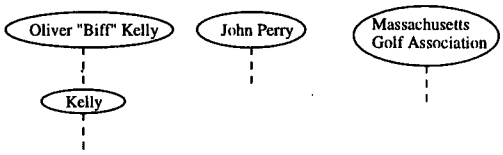

Figure 5: Coreference Chains for doc.38

that contains "John Perry"), the Sentence Extractor module extracts all the sentences that contain the noun phrases which form the coreference chain. In other words, the SentenceExtractor module produces a "summary" of the article with respect to the entity of interest. These summaries are a special case of the query sensitive techniques being developed at Penn using CAMP. Therefore, for doc.36 (Figure 2), since at least one of the three noun phrases ("John Perry," "he," and "Perry") in the coreference chain of interest appears in each of the three sentences in the extract, the summary produced by SentenceExtractor is the extract itself. On the other hand, the summary produced by SentenceExtractor for the coreference chain of interest in doc.38 is only the first sentence of the extract because the only element of the coreference chain appears in this sentence.
- Finally, for each article, the VSM-Disambiguate module uses the summary extracted by the SentenceExtractor and computes its similarity with the summaries extracted from each of the other articles. The VSM-Disambiguate module uses a standard vector space model (used widely in information retrieval) (Salton, 89) to compute the similarities between the summaries. Summaries having similarity above a certain threshold are considered to be regarding the same entity.

4.1 Scoring

In order to score the cross-document coreference chains output by the system, we had to map the cross-document coreference scoring problem to a within-document coreference scoring problem. This was done by creating a meta document consisting of the file names of each of the documents that the system was run on. Assuming that each of the doc-

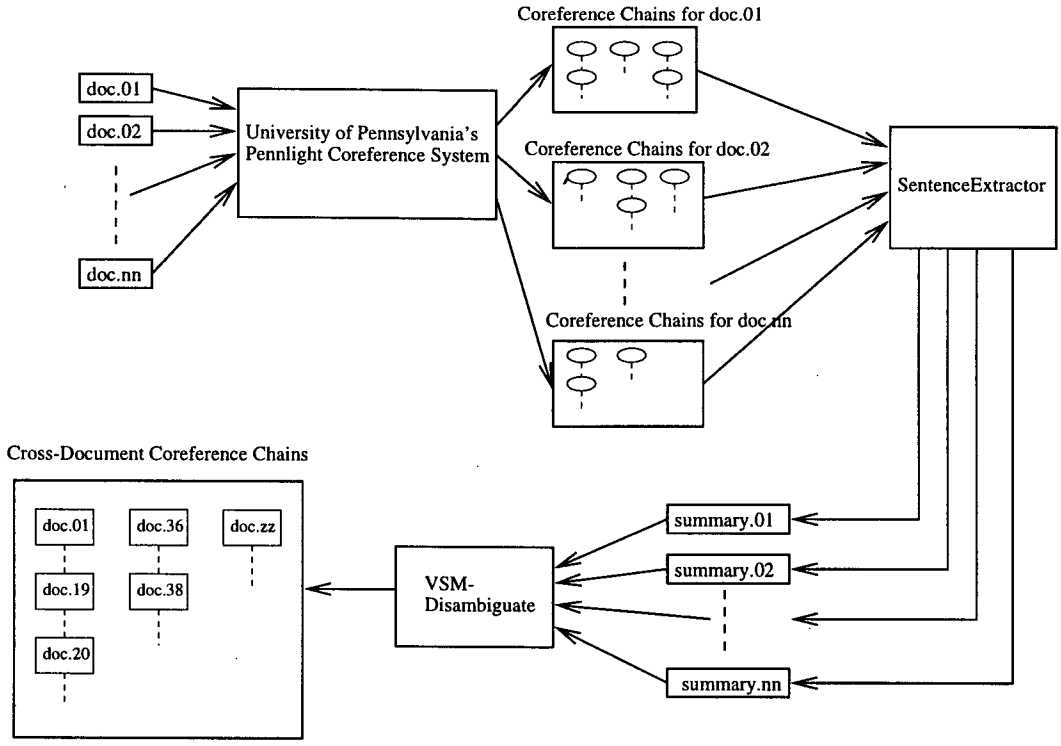

Figure 1: Architecture of the Cross-Document Coreference System

uments in the data sets was about a single entity, or about a single event, the cross-document coreference chains produced by the system could now be evaluated by scoring the corresponding within-document coreference chains in the meta document.

We used two different scoring algorithms for scoring the output. The first was the standard algorithm for within-document coreference chains which was used for the evaluation of the systems participating in the MUC-6 and the MUC-7 coreference tasks. This algorithm computes precision and recall statistics by looking at the number of links identified by a system compared to the links in an answer key.

The shortcomings of the MUC scoring algorithm when used for the cross-document coreference task forced us to develop a second algorithm - the B-CUBED algorithm - which is described in detail below. Full details about both these algorithms (including the shortcoming of the MUC scoring algorithm) can be found in (Bagga, 98).

4.1.1 The B-CUBED Algorithm

For an entity, i, we define the precision and recall with respect to that entity in Figure 6.

The final precision and recall numbers are computed by the following two formulae:

$$\text{Final Precision} = \sum_{i=1}^{N} w_i * Precision_i$$

$$\text{Final Recall} = \sum_{i=1}^{N} w_i * Recall_i$$

where N is the number of entities in the document, and w_i is the weight assigned to entity i in the document. For the results discussed in this paper, equal weights were assigned to each entity in the meta document. In other words, $w_i = \frac{1}{N}$ for all i.

5 Cross-Document Coreference for Events

In order to extend our systems, as described earlier, so that it was able to handle events, we needed to figure out a method to capture all the information about an event in a document. Previously, with named entities, it was possible to use the within-document coreference chain regarding the entity to extract a "summary" with respect to that entity. However, since CAMP does not annotate within-document coreference chains for events, it was not possible to use the same approach.

The updated version of the system builds "summaries" with respect to the event of interest by extracting all the sentences in the article that contain either the verb describing the event or one of its nominalizations. Currently, sentences that contain synonyms of the verb are not extracted. However, we did conduct an experiment (described later in the paper) where the system extracted sentences con-

taining one of three pre-specified synonyms to the verb.

The new version of the system was tested on several data sets.

5.1 Analysis of Data

Figure 7 gives some insight into the data sets used for the experiments described later in the paper. In the figure, Column 1 shows the number of articles in the data set. The second column shows the average number of sentences in the summary for the entity/event of interest constructed for each article. Column 3 shows, for each summary, the average number of words that were found in at least one other summary (in the same data set). The conditions when measuring the overlap should be noted here:

- the summaries are filtered for stop words
- all within-document coreference chains passing through the summaries are expanded and the resulting additional noun phrases are attached to the summaries

The fourth column shows for each such overlapping word, the average number of summaries (in the same data set) that it is found in. Column 5 which is the product of the numbers in Columns 3 and 4 shows, for each summary, the average number of summaries, in the data set, it shares a word with (the amount of overlap). We hypothesize here that the higher the amount of overlap, the higher is the ambiguity in the domain. We will return to this hypothesis later in the paper.

Figure 7 shows that the "resign" and the "espionage" data sets are remarkably similar. They have very similar numbers for the number of sentences per summary, the average number of overlapping words per summary, and the average number of summaries that each of the overlapping words occur in. A closer look at several of the summaries from each data set yielded the following properties that the two data sets shared:

- The summaries usually consisted of a single sentence from the article.
- The "players" involved in the events (people, places, companies, positions, etc.) were usually referenced in the sentences which were in the summaries.

However, the "election" data set is very different from the other two sets. This data set has almost twice as many sentences per summary (2.38). In addition, the number of overlapping words in each summary is also comparatively high although the average number of summaries that an overlapping words occurs in is similar to that of the other two data sets. But, "elections" has a very high overlap

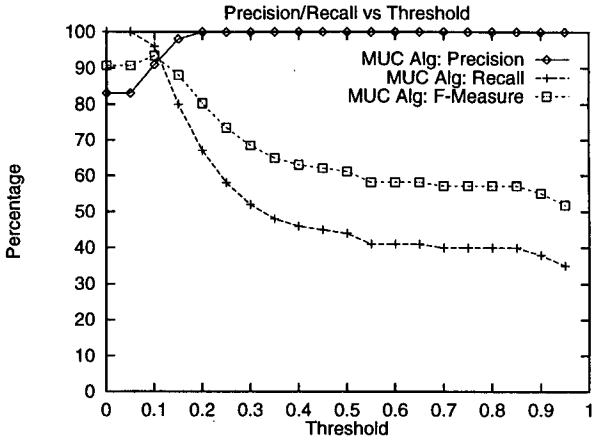

Figure 8: Results for the "John Smith" data set using the MUC scorer

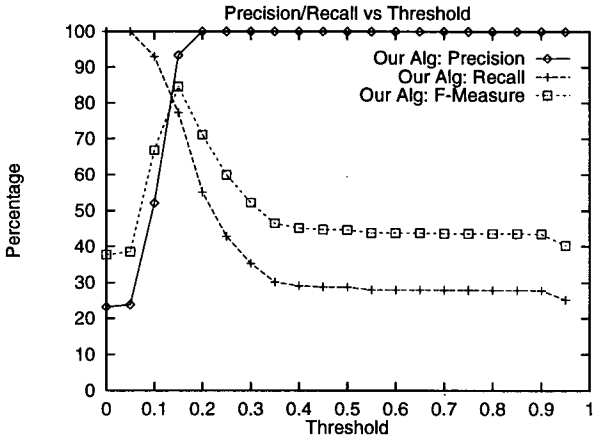

Figure 9: Results for the "John Smith" data set using the B-CUBED scorer

number (22.41) which is about 30% more than the other data sets. From our hypothesis it follows that this data set is comparatively much more ambiguous; a fact which is verified later in the paper.

Assuming our hypothesis is true, the overlap number also gives an indication of the optimal threshold which, when chosen, will result in the best precision and recall numbers for the data set. It seems a reasonable conjecture that the optimal threshold varies inversely with the overlap number i.e. the higher the overlap number, the higher the ambiguity, and lower the optimal threshold.

5.2 Experiments and Results

We tested our cross-document coreference system on several data sets. The goal was to identify cross-document coreference chains about the same event.

Figures 8 – 15 shows the results from the experiments we conducted. For each experiment conducted, the following conditions hold:

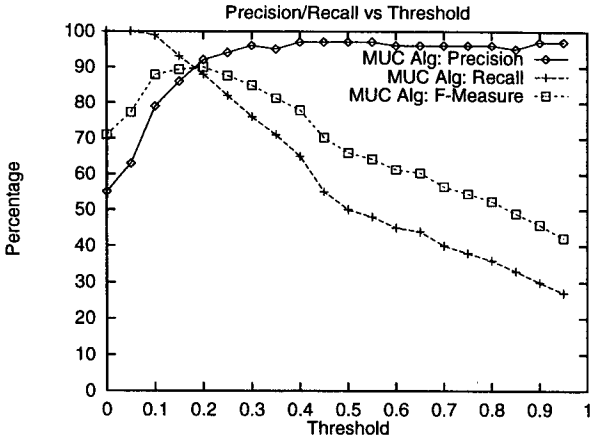

Figure 10: Results for the "resign" data set using the MUC scorer

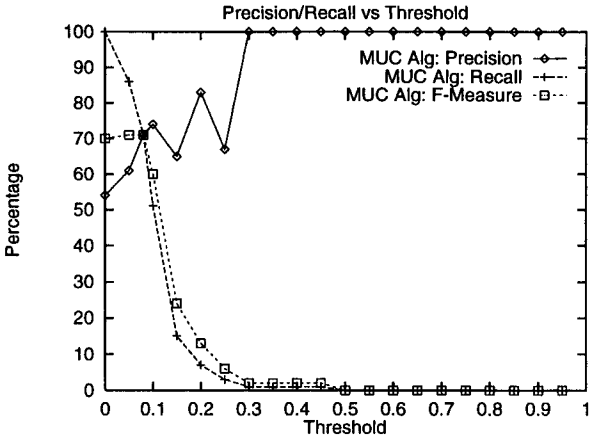

Figure 12: Results for the "elections" data set using the MUC scorer

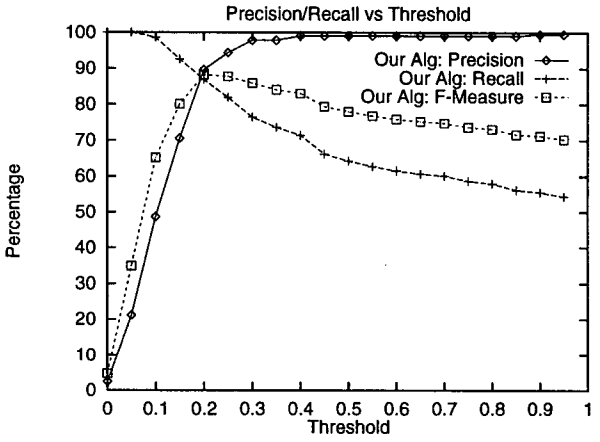

Figure 11: Results for the "resign" data set using the B-CUBED scorer

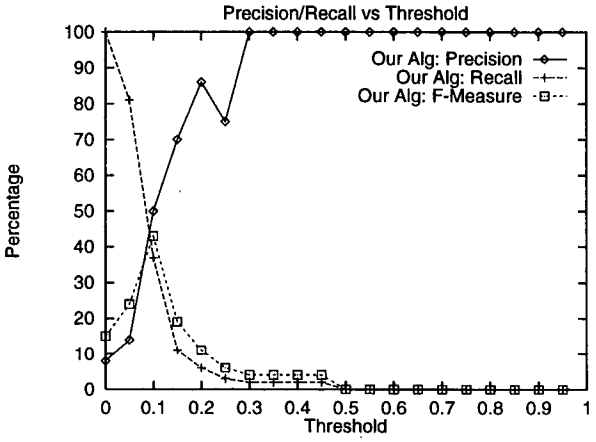

Figure 13: Results for the "elections" data set using the B-CUBED scorer

- Figure 7 shows, for each data set, the number of articles chosen for the experiment.
- All of the articles in the data sets were chosen randomly from the 1996 and 1997 editions of the New York Times. The sole criterion used when choosing an article was the presence/ absence of the event of interest in the data set. For example, an article containing the word "election" would be put in the elections data set.
- The answer keys for each data set were constructed manually, although scoring was automated.

Figure 16 shows for each data set, the optimal threshold, and the best precision, recall, and F-Measure obtained at that threshold.

5.3 Analysis of Results

We had mentioned earlier that we expected the optimal threshold value to vary inversely with the overlap number. Figure 16 verifies this - the optimal thresholds decline for the "espionage", "resign", and the "election" data sets (which have increasing overlap numbers). In addition, the results for the "election" data set also verify our hypothesis that data sets with large overlap numbers are more ambiguous.

There are several different factors which can affect the performance of the system. We describe some of the more important ones below.

expansion of coreference chains: Expanding the coreference chains that pass through the sentences contained in a summary and appending the coreferent noun phrases to the summary results in approximately a 5 point increase in F-Measure for each data set.

use of synonyms: For the "election" data set, the use of three synonyms (poll, vote, and campaign) to extract additional sentences for the summaries helped in increasing the performance

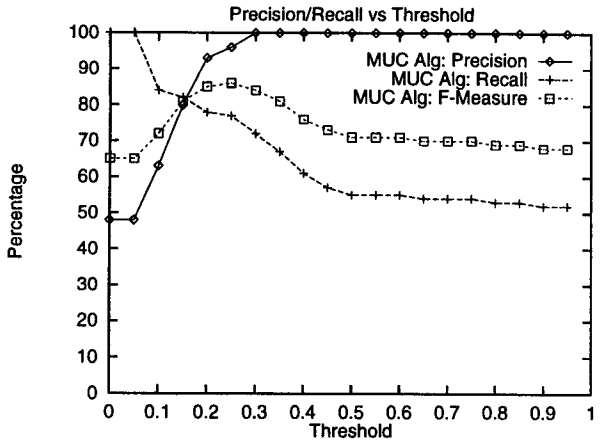

Figure 14: Results for the "espionage" data set using the MUC scorer

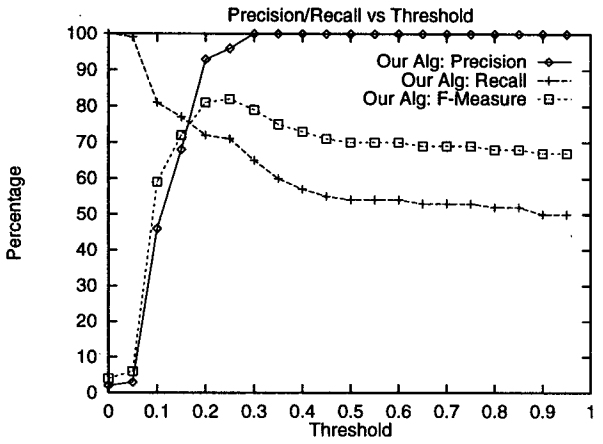

Figure 15: Results for the "espionage" data set using the B-CUBED scorer

of the system by 3 F-measure points. The resulting increase in performance implies that the sentences containing the term "election" did not contain sufficient information for disambiguating all the elections. Some of the disambiguation information (example: the "players" involved in the event) was mentioned in the additional sentences. This also strengthens our observation that this data set is more comparatively more ambiguous.

presence of a single, large coreference chain: The presence of a single, large cross-document coreference chain in the test set affects the performance of a system with respect to the scoring algorithm used. For example, the "election" data set consisted of a very large coreference chain - the coreference chain consisting of articles regarding the 1996 US General (Congressional and Presidential) elections. This chain consisted of 36 of the 73 links in the data set. The B-CUBED algorithm penalizes systems severely for precision and recall errors in such a scenario. The difference in the results reported by the two scoring algorithms for this data set is glaring. The MUC scorer reports a 71 point F-Measure while the B-CUBED scorer reports only a 43 point F-Measure.

5.4 The "election" Data Set

Since the results for the "election" data set were significantly lower than other results, we decided to analyze this data set in more detail. The following factors makes this data set harder to deal with:

presence of sub-events: The presence of sub-events that correspond to a single event makes the task harder. The "election" data set often mentioned election events which consisted of more than one actual election. For example, the data set contained articles which mentioned the 1996 US General Elections which comprised of the US Congressional elections and the US Presidential elections. In addition, there were articles which only mentioned the sub-elections without mentioning the more general event.

"players" are the same: Elections is one event where the players involved are often the same. For example, elections are about the same positions, in the same places, and very often involving the same people making the task very ambiguous. Very often the only disambiguating factor is the year (temporal information) of the election and this too has to be inferred. For example, articles will mention an election in the following ways: "the upcoming November elections," "next years elections," "last fall's elections," etc.

descriptions are very similar: Another very important factor that makes the "elections" task harder is the fact that most election issues (across elections in different countries) are very similar. For example: crime rates, inflation, unemployment, etc.

6 Interannotator Agreement

When comparing machine performance against a human annotation, it is important to understand how consistently two humans can perform the same task. If people cannot replicate one other, then there may be serious problems with the task definition that question the wisdom of developing automated methods for the task.

Both authors independently annotated the "elections" data set with no agreed upon annotation standard in contrast to how data sets were annotated in the MUC-6/7 coreference task. Instead, we used

whatever mutual understanding we had on what the goal of our annotation was from phone calls over the course of a few months. We did not develop an annotation standard because we have not considered a sufficiently broad range of events to write down necessary and sufficient conditions for event coreference. For now our understanding is:

Any two events are in the same equivalence class if they are of the same generic class, ie "elections" or "resignations", and the principle actors, entities, and times are the same.

This definition does not cover the specificity of event descriptions, i.e. the difference between the general November 96 elections and a particular election in a district (at the same time). We left this decision up to human judgment rather than trying to codify the decision at this early stage.

Interannotator agreement was evaluated in two phases, a completely independent phase and a consensus phase where we compared annotations and corrected obvious errors and attentional lapses but allowed differences of opinion when there was room for judgment. The results for the completely independent annotation were 87% precision and 87% recall as determined by treating one annotation as truth and the other as a systems output with the MUC scorer. Perfect agreement between the annotators would result in 100% precision and recall. These results are quite high given the lack of a clear annotation standard in combination with the ambiguity of the task.

After adjudication, the agreement increased significantly to 95% precision and recall which indicates that there was genuine disagreement for 5% of the links found across two annotators. Using the B-CUBED scorer the results were 80% for the independent case and 93% for the consensus phase. These figures establish an upper bound on possible machine performance and suggest that cross document event coreference is a fairly natural phenomenon for people to recognize.

7 Future Research

The goal of this research has been to gain experience in cross document reference across a range of entities/events. We have focused on simple techniques (the vector space model) over rich data structures (within document coreference annotated text) as a means to better understanding of where to further explore the phenomenon.

It is worth exploring alternatives to the vector space model since there are areas where it could be improved. One possibility would be to explicitly identify the individuating factors of events, i.e. the "players" of an event, and then individuate by comparing these factors. This would be particularly helpful when there is only one individuating factor like a date that differentiates two events.

The benefit of cross document entity reference centers around novel interfaces to large data collections, so we are focusing on potential applications that include link visualization (Bagga, 98c), question answering, and multi-document summarization.

8 Conclusions

We have shown that it is possible to extend our earlier work with cross document person reference to include cross document event reference. This is achieved by using the vector space model to form equivalence classes of "summaries" about the events in question. These summaries are generated by including sentences that have coreference into the core event sentence as well as sentences that fit within a synonymy class for the event in question. Our results are encouraging with performance ranging from 45% f-score to 90% f-score. We also have established that human annotators agree on cross document event reference around 95% of the time.

References

Bagga, Amit, and Breck Baldwin. Algorithms for Scoring Coreference Chains. *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation*, May 1998.

Bagga, Amit, and Breck Baldwin. How Much Processing is Required for Cross-Document Coreference? *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation*, May 1998.

Bagga, Amit, and Breck Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (COLING-ACL'98), pp. 79-85, August 1998.

Bagga, Amit, and Breck Baldwin. Coreference as the Foundations for Link Analysis Over Free Text Databases. In *Proceedings of the COLING-ACL'98 Content Visualization and Intermedia Representations Workshop* (CVIR'98), pp. 19-24, August 1998.

Grishman, Ralph. Whither Written Language Evaluation?, *Proceedings of the Human Language Technology Workshop*, pp. 120-125, March 1994, San Francisco: Morgan Kaufmann.

Salton, Gerard. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, 1989, Reading, MA: Addison-Wesley.

$$\text{Precision}_i = \frac{\textit{number of correct elements in the output chain containing entity}_i}{\textit{number of elements in the output chain containing entity}_i}$$

$$\text{Recall}_i = \frac{\textit{number of correct elements in the output chain containing entity}_i}{\textit{number of elements in the truth chain containing entity}_i}$$

Figure 6: Definitions for Precision and Recall for an Entity i

data set	# of articles	avg # of sentences per summary	avg # of overlapping words in summary	avg # of summaries that overlapping words occur in	amount of overlap per summary
John Smith	197	1.16	2.46	5.74	14.13
resign	219	1.35	4.35	3.99	17.36
elections	135	2.38	5.66	3.96	22.41
espionage	184	1.28	4.57	3.62	16.54

Figure 7: Analysis of the Data Sets

Data Set	Scorer	Optimal Threshold	F-Measure	Precision	Recall
John Smith	MUC	0.15	88	98	80
	B-CUBED	0.15	84.6	93.3	77.3
resign	MUC	0.20	90	92	88
	B-CUBED	0.20	88.2	89.6	86.8
elections	MUC	0.08	71	71	71
	B-CUBED	0.10	43	50	37
espionage	MUC	0.25	86	96	77
	B-CUBED	0.25	82	96	71

Figure 16: Analysis of the Data Sets