UNIVERSITY OF PENNSYLVANIA: DESCRIPTION OF THE UNIVERSITY OF PENNSYLVANIA SYSTEM USED FOR MUC-6

Breck Baldwin
Jeff Reynar
Mike Collins
Jason Eisner
Adwait Ratnaparkhi
Joseph Rosenzweig
Anoop Sarkar
Srinivas

University of Pennsylvania
Department of Computer and Information Sciences
200 South 33rd Street
Philadelphia, PA 19104

breck@linc.cis.upenn.edu, jcreynar@linc.cis.upenn.edu

(215) 898-0326

INTRODUCTION

Breck Baldwin and Jeff Reynar informally began the University of Pennsylvania's MUC-6 coreference effort in January of 1995. For the first few months, tools were built and the system was extended at weekly 'hack sessions.' As more people began attending these meetings and contributing to the project, it grew to include eight graduate students. While the effort was still informal, Mark Wasson, from Lexis-Nexis, became an advisor to the project. In July, the students proposed to the faculty that we formally participate in the coreference task. By that time, we had developed some of the system's infrastructure and had implemented a simplistic coreference resolution system which resolved proper nouns by means of string matching. After much convincing, the faculty agreed at the end of July that we could formally participate in MUC-6. We then began an intensive effort with full-time participation from Baldwin and Reynar, and part-time efforts from the other authors. In August we were given permission from Yael Ravin of IBM's Information Retrieval group to use the IBM Name Extraction Module [3]. We were also given access to a large acronym dictionary which Peter Flynn maintains for a world wide web site in Iceland (http://curia.ucc.ie/info/net/acronyms/acro.html).

The vast majority of our system was developed in August and September. Our efforts prior to that time were mostly directed towards implementing a parallel-file data structure which allowed new components to be added quickly with minimal effort. The ease of incorporating new components was demonstrated by the addition of a full syntactic parser two weeks prior to the evaluation. In this data structure, enhancements to the input data, such as tokenization, part-of-speech tags, or parse trees, are stored in separate, aligned files. As a result, building a new module which requires input from earlier components is as simple as loading the files created by those components and performing the necessary processing. The fact that modules further along in the pipeline do not alter the output of earlier components means that output files can be read-only. As a result, the system is afforded a measure of robustness: if one component fails, further components will not necessarily be crippled and no downstream component can alter the output of an earlier component.

This simple data structure, which was inspired by a pretty-printing convention used by Lexis-Nexis to display multiple levels of textual annotation, also allowed people to write software in the programming language of their choice. Ultimately, the majority of the code written explicitly for MUC was in Perl 4, but some programs were also written in C and several different shell languages. Other system components not developed explicitly for MUC were written in Lisp and C++.

Despite the advantages of this approach, the parallel-file data structure had some drawbacks. First, because the system was built using many small tools, the number of files grew to be quite large, nearly 100 per article. As a result, disk space became a problem. Second, because of the large number of files and the number of processes reading each of them, file access time accounted for a significant portion of the time required to run the system. It took approximately 12 minutes to process an average length article when processing was done in batch mode. Processing input files in groups allowed the overhead for loading dictionaries and statistical models to be reduced because it could be averaged over many articles.

Our coreference resolution system was built from several components, each of which addressed different types of coreference. The philosophy behind this methodology was that high precision components could be linked together serially to build an easily extensible, modular system. We focused on building high precision components on the assumption that many high precision, moderate recall components, when linked together, would yield a system with good overall recall. This goal was met with varying degrees of success. Unfortunately, only one of the three components which posited coreference emerged as being highly precise: the proper name matching component.

We utilized off-the-shelf components whenever possible. Most of these tools were developed at Penn. As a result, the majority of our efforts went into writing parsers and preprocessing utilities which allowed various pre-existing tools to communicate with one another and produce output which could be used by other tools further in the processing pipeline. Thus, we were freed to spend time developing the task-specific components of the system and performing data analysis. Although no time was spent developing tools particularly for the MUC task prior to January, many hours went into developing some of the off-the-shelf components we used, such as Eric Brill's part-of-speech tagger [2] and Lance Ramshaw and Mitch Marcus' Noun Phrase Detector [10]. We estimate the total number of hours spent on the project itself to be roughly 1800, distributed among the eight graduate students who worked on the project. The vast majority of these hours were contributed between the end of July and competition week in early October.

Table 1 shows the performance of our system when simple formatting errors, which hurt performance on two of the 30 test files, were corrected. Table 2 contains our official system performance figures. Table 3 contains system performance when optional elements were treated as if required. This set of scores is presented in order to allow comparison between scores for various system components without having to deal with the adjustment to the number of correct items which results from different components marking coreference between different numbers of optional elements.

Recall	973/1540	.63
Precision	973/1345	.72

Table 1: System Performance without Formatting Errors.

Recall	848/1529	.55
Precision	848/1345	.63

Table 2: Official System Performance.

Recall	973/1627	.60

Precision	973/1345	.72

Table 3: System Performance without Formatting Errors but with optional elements treated as required.

THE SYSTEM

Throughout the system description section, words and phrases which appear in articles will be displayed in *italics*. Figure 1 contains a system flowchart. Databases are shown in drums and system modules are shown in rectangles.

End of Sentence Detection

The first step in our processing pipeline is end-of-sentence detection. Sentence boundaries are identified using a maximum entropy model developed explicitly for MUC-6. This model was built quickly using a general maximum entropy modeling tool which will be discussed in a forthcoming paper [11]. Sentence final punctuation is defined to include only periods, exclamation points and question marks; we do not attempt to mark sentence boundaries indicated by semi-colons, commas or conjunctions. Only instances of sentence-final punctuation which are immediately followed by white space or symbols which may legitimately follow sentence boundaries, such as quotation marks, were considered to be potential sentence boundaries. For convenience, we define any sequence of white-space separated tokens to be a word while discussing this stage of processing.

The maximum entropy model was trained using the dry run and training portions of the MUC-6 coreference annotated data, which included SGML annotated sentence boundaries. The model used binary-valued features of the word to which the putative end-of-sentence marker was conjoined, as well as binary-valued features of the preceding and following words. These features included whether the word was a corporate designator, such as *Corp.* or *Inc.*, or an honorific, such as *Dr.* or *Ms.*; whether the word was upper-case; whether the word was a likely monetary value; whether the word was likely to be a percentage; whether the word was a number; whether the word contained punctuation indicative of ellipsis; and features indicating whether the word ended in various non-alphanumeric characters.

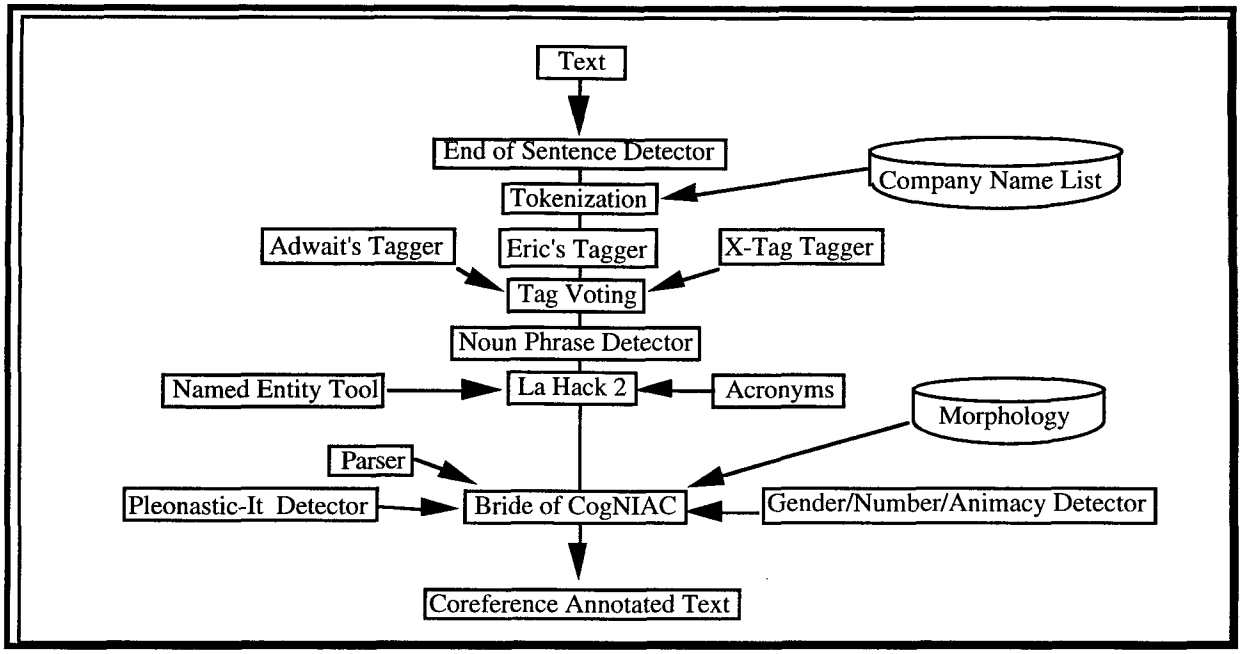

Figure 1: System flowchart.

We did not subject this component to rigorous testing, but did examine its output for approximately 300 blind test sentences and found that only one error was made. We intend to further refine this component and subject it to automatic testing against a sentence-detected corpus in the near future.

Tokenization

Once sentence boundaries are identified, tokenization begins. We developed our tokenizer solely for the MUC coreference task because of specific tokenization requirements. The combination of the character-based nature of the scoring software and the requirements of various tools that punctuation be separated from words forced us to build a tokenizer which maintains a character offset mapping for all of the tokens in the input messages. A trivial error in this system caused two of the 30 test messages to be garbled sufficiently that the scorer detected virtually no correct coreference in them. This is why we are presenting both official and unofficial scores.

In addition to maintaining the character offset mapping, the tokenizer performs four non-standard tasks. The first is the alteration of headline word capitalization. The Wall Street Journal adheres to standard conventions for capitalization of words in headlines, but since capitalization is an important cue for coreference resolution, we attempted to eliminate capitalization which resulted solely from these conventions. Headline words which were capitalized in the body of the text anywhere other than sentence-initial position remained capitalized, as did those which were frequently capitalized other than in sentence-initial position in the Treebank Wall Street Journal corpus [8]. All other uppercase words were converted to lowercase.

The second non-standard task addressed by the tokenizer is the extraction of date information. The dateline field is parsed to determine when each article was written. This information is later used to posit coreference between words or phrases such as *today*, *tomorrow*, *this week*, *this year*, and dates, such as *November 20, 1995*.

The third non-standard component determines whether *'s* or *'* is a genitive marker or part of a company name. When it is actually part of a company name, it does not indicate possession of the following noun phrase. This step was necessary because the part-of-speech taggers and the noun phrase detector required genitive markers to be tokenized separately, while non-genitive instances of *'s* or *'* were required to remain attached. For instance,

McDonald's, when it refers to the fast-food chain, should be treated as a single token, while *Mary's* should be separated into two tokens: *Mary* and *'s*.

The final unique task the tokenizer addresses is hyphenated-word splitting. Since coreference is allowed between portions of hyphenated words which are themselves words, such as *Apple* in the phrase *a joint Apple-IBM venture*, determining whether a portion of a hyphenated word may participate in coreference is important. The heuristic we use is similar to the one used to determine whether a headline word should be downcased. That is, when one or more of the words which comprise a hyphenated word exists on their own within the article, then the hyphenated word is split into multiple tokens.

Unfortunately, because of the nature of the training data used by the noun phrase detector, bare hyphens cause serious noun phrase detection errors. For simplicity, and because of time limitations, we opted not to retrain the noun phrase detector. As a result, multiple tokenizations of each article are maintained. In one of the tokenizations, hyphenated words are left unaltered. In the other version, hyphenated words are split into multiple tokens based on the above criteria. Also, the tokenizer is responsible for maintaining the mapping between these two tokenizations so that the output of tools which use different tokenization schemes can be combined.

Part-of-Speech Tagging

Several components of the MUC coreference system, such as the noun phrase detector, require part-of-speech (POS) tags for all of the words in an article. We combined the output of the following three POS taggers using a simple voting scheme: Eric Brill's Rule Based Tagger version 1.14 [2], the XTAG tagger, which is an implementation of Ken Church's PARTS tagger [4] and Adwait Ratnaparkhi's Maximum Entropy Tagger [11]. Each of these taggers uses the Penn Treebank tagset [8].

These three taggers, which were trained on the Penn Treebank Wall Street Journal corpus, tag pre-tokenized text. The tag actually used by the MUC system is determined by a majority voting scheme, in which a tag is chosen as the "winner" if at least two of the taggers postulate it. In the rare event that all three taggers disagree, the system uses the tag assigned by the maximum entropy tagger. In most cases, the majority voting scheme eliminates errors that are esoteric to a single tagger, and should therefore perform better than any single tagger. We did not have time to empirically verify this hypothesis, but intend to do so in the future. We may also improve upon the voting model by incorporating information regarding which tagger proposed each tag.

Basal Noun Phrase Detection

To identify noun phrases, the system uses Lance Ramshaw and Mitch Marcus' basal noun phrase detector [10]. Basal noun phrases are those noun phrases in the lowest level of embedding in the Penn Treebank's annotations. Intuitively, they are the smallest noun phrases in a parse. For example, *chief executive officer* and *International Business Machines* are both basal noun phrases, but *chief executive officer of International Business Machines* is not, since it contains nested noun phrases. Ramshaw and Marcus' noun phrase detector is based on Eric Brill's work on learning transformational rules for part-of-speech tagging. It was trained using a section of the tagged and parsed Treebank Wall Street Journal corpus disjoint from the MUC-6 test data.

We postprocess the output of their tool to make it more appropriate for the coreference task. For instance, it brackets noun phrases containing genitives in the following way: [Noun Phrase 1] ['s Noun Phrase 2]. But, we prefer [Noun Phrase 1] 's [Noun Phrase 2] since it is more appropriate for further processing steps. In addition, we manually added some transformations to the set learned from the treebank. These transformations generalized on learned ones. For instance, rules were learned which involved days of the week, but due to sparsity of training data, they were learned only for a subset of the seven days of the week. We manually added the missing cases. We did not independently measure the performance of their tool using this modified rule set, but may do so in the future.

Knowledge Sources

We experimented with various knowledge sources during system development, including WordNet [9], the XTAG morphological analyzer [6], Roget's publicly available 1911 thesaurus, the Collins dictionary, a version of the American Heritage dictionary for which the University of Pennsylvania has a site license and the Gazetteer. Only WordNet, the XTAG morphological analyzer and the Gazetteer were used in the final system.

We extracted a geographic name database from a publicly available version of the Gazetteer which we downloaded from the Center for Lexical Research. This database contains names of continents, islands, island groups, countries, provinces, cities and airports. This information is used when performing type checking prior to positing coreference between entities.

The XTAG morphology database [6] was originally extracted from the 1979 edition of the Collins English Dictionary and the Oxford Advanced Learner's Dictionary of Current English, and then edited and augmented by hand. It contains approximately 317,000 inflected items, along with their root forms and inflectional information, such as case, number and tense. Thirteen parts of speech are differentiated: noun, proper noun, pronoun, verb, verb particle, adverb, adjective, preposition, complementizer, determiner, conjunction, interjection, and noun/verb contraction. Nouns and verbs are the largest categories, with approximately 213,000 and 46,500 inflected forms, respectively.

Tagging for Gender, Number and Animacy

To resolve pronouns which typically select for a gendered antecedent as well as those that typically select for an animate antecedent, gendered or non-gendered, the WordNet 1.5 lexical database [9] for nouns is used to tag each potential antecedent with respect to these semantic features. In addition, rudimentary morphological analysis of the head of a noun phrase is performed and several databases are consulted to determine whether a particular noun phrase refers to a male, a female, or a person of either gender. Also, some singular count nouns, such as *committee*, may be the antecedents of plural pronouns. WordNet is also consulted to tag such nouns as possibly having sets of individuals as their referent.

WordNet's noun database is organized as an inheritance lattice. For example, the entry for *man* is linked to daughter nodes which include the entries *bachelor*, *boyfriend*, *eunuch*, etc. Assuming that a semantic feature such as maleness generally will propagate from a parent in the hierarchy to its children, one can test the gender of a given noun by examining its ancestors. If one of the ancestors is the entry *male*, for example, it may be concluded that the word itself typically denotes an entity which is male. Similarly, the WordNet entry *social_group* tends to subsume nouns which can have groups of individuals as their referents.

Unfortunately, the WordNet taxonomy is more like a tree than a lattice, so that many useful multiple inheritance links do not exist. For example, the entry for *uncle* is not a descendant of the entry for *man*, although an uncle is clearly a type of man. Additionally, as with any semantic inheritance hierarchy, not all features are always passed down from parent to child, so that strictly monotonic reasoning is not valid.

To ameliorate these deficiencies and complications, the query to WordNet takes the form of a Boolean query about the ancestors of a given word entry. For example, an OR operator is used to tag as male words which are descendants of either the *male* node or the *kinsman* node, which subsumes *uncle*. This supplants the missing inheritance link, which would be needed in a complete semantic taxonomy, between *male* and *kinsman*. To prune out descendants of an entry such as *man* which do not inherit the semantic feature of maleness, an AND NOT operator can be used to exclude subclasses of the class of descendants of *male*. Additionally, to circumvent problems with solely relying on the Boolean query, a word's definition is also examined in a rudimentary way, to check for key words that indicate semantic features of the potential referents of this word, such as the word *someone*, which suggests a human referent.

For polysemous words, WordNet may give conflicting evidence because of the word's multiple senses. For example, *end* is judged as potentially compatible with a human referent, because an end is a type of football player. But in most contexts, this sense of *end* will be wrong and this word should not be considered as the potential antecedent for a pronoun such as *he*.

Therefore, the evidence from WordNet is weighted on a scale of plausibility. The evidence for *uncle* is considered more plausible than that for *end* because both senses of *uncle* in WordNet have the entry *person* among their ancestors. On the other hand, only one of the thirteen senses for *end* has *person* as its ancestor. Moreover, not all of the senses of *end* are equally likely to occur. The WordNet semantic concordance provides frequency information from a fraction of the Brown Corpus for senses of *end* and other words in the noun database. These counts can be used to estimate the probabilities of WordNet word senses. When no data is available from the semantic concordance for some senses of a word, the gaps in frequency are smoothed. If no data is available for any sense of the word, the uniform distribution is assumed.

The evidence from WordNet is then weighted according to how likely it is that the sense for which the evidence is obtained is the correct sense of the word seen in the input file. A more sophisticated approach would involve using word-sense disambiguation techniques to guess the correct sense of the word, and then only query WordNet about that particular sense. However, the method employed in the current system is able to discriminate reliably on a coarse level between cases like *end* and *uncle*. A weight of 1.0 is assigned to the person feature for *uncle*, whereas only 0.024 is assigned to this feature in *end*.

As a second source of evidence about the gender or animacy of noun phrase referents, two tables of gendered first names, compiled by Mark Kantrowitz and Bill Ross and freely available from the Computing Research Laboratory of New Mexico State University, are consulted. The table of first names overlaps with place names and time words. For example, *Canada* and *Tuesday* are women's names. In such cases, the evidence from the table is discarded. This evidence is weighted separately from the WordNet look-up results.

Finally, a rough analysis of the suffix morphology of the word is undertaken. Nouns ending in "-man" which do not end in "-woman" tend to denote male humans. However, due to the inherent gender bias of language, words such as *chairman* can also be used to refer to women. Hence such words also count as evidence of a female referent, but to a lesser degree. This results in both the male weight and the female weight being set to non-zero values. The difference in weighting between the two is currently based on intuition, though corpus methods might yield a more exact estimate of how much weight to give the female reading based on how often such words are actually used to refer to women.

Pleonastic It Detection

It is often used anaphorically in Wall Street Journal Text. Nonetheless, identifying instances of pleonastic it, which do not corefer, is still significant. The system identifies these instances of *it* by scanning tagged text and applying partly syntactic and partly lexical tests. Most of these tests are described in [7], but some additional tests were added to increase coverage. The fifteen rules used to detect pleonastic it are shown below in table 4. Part of speech tags follow words and a slash, and are specified using the Penn Treebank tagset. Disjunctions are indicated using a vertical bar, (|), and optional elements are surrounded by brackets, ([]). S abbreviates sentence; NP means noun phrase; and VP stands for verb phrase. We abbreviate CA for comparative adjectives, such as *larger* or *smaller*; SA for superlatives, such as *greatest* or *largest*; MA for modal adjectives, such as *necessary* or *uncertain*; MV for modal verbs, like *could* or *will*; CV for cognitive verbs, such as *recommended* or *hoped*; and CADV and SADV for comparative and superlative adverbs.

It is (CA/JJR | SA/JJR | not) MA/JJ that S	It (is not | may be) (CA/JJR | SA/JJR | not) MA/JJ ...
I MV appreciate|believe it if ...	It MV be (MA/JJ | CV/VBD) ...
It is (CA/JJR | SA/JJR | not) CV/VBD that S	It (seems|appears|means|follows) [that] S
... NP makes|finds it MA/JJ [for NP] to VP ...	It is time to VP ...
It is thanks to NP that S	It is (CADV/RB | SADV/RB) adj/JJ ...
It (signals|is/VBZ) ?/NNP ?/POS ?/NN ...	... (makes | made) it clear that S
It is a (CADV/RB | SADV/RB) MA/JJ NP ...	Would n't it be (CA/JJR | SA/JJR | not) MA/JJ ...
It is (CA/JJR | SA/JJR | not) MA/JJ [for NP] to VP ...	

Table 4: Pleonastic it detection rules.

La Hack 2

The first component of the system which actually marks coreference between entities is called La Hack 2. It's performance is shown in table 5. Our first attempt at a coreference system, La Hack 1, posited coreference between identical upper case words in the text, and was written to test the validity of the system's SGML annotation and to test the tokenizer. La Hack 2 was written to do more sophisticated string matching. It uses several knowledge sources, including the IBM Name Extraction Module, and a simple unification system to produce coreference chains. The knowledge sources are used to determine whether an entity is of type person, place, corporation or other. Most of the entities which La Hack 2 annotates are proper nouns, but the date information extracted by the tokenizer is used here as well. The majority of the strings annotated are noun phrases detected by the noun phrase detector, but some sub-noun phrase units are annotated as well. Proper nouns which are portions of longer noun phrases may be annotated. For example, *Apple* in the phrase *Apple stock prices* would be annotated if there were other references to *Apple* in the article.

La Hack 2 makes four passes through each article. On the first, it builds coreference chains containing alternate forms of corporate and person names as identified by the Name Extraction Module. These variant references include references to people by first name only, last name only, last name and an honorific, and references which omit middle names. For instance, *General Colin Powell* could be referred to as *General Powell*, *Colin*, *Powell*, *Mr. Powell* and so forth. Variant corporate names may be references which exclude corporate designators, use acronyms or omit a company's industry. For example, *Apple Computer Inc.* might be referred to as *Apple*, *Apple Inc.*, etc.

The next processing step looks for date matches, and those alternate forms not identified by the IBM tool. The third step looks for upper case string matches which are not variant name references or which do not contain corporate designators or honorifics. Product names, some acronyms and miscellaneous other upper case words are entered into coreference chains in this stage. The final stage is an upper case substring match which is targeted at finding coreference chains which were missed by the named entity tool and the other stages as well.

The purpose of the simple type system is mainly to prevent coreference chains from being created by the substring matching stage which contain substrings of different types. For instance, *Apple* is a substring of *Apple CEO John Sculley*, but they cannot be coreferent since *John Sculley* is a person and *Apple* is a corporation.

Recall	468/1627	29%
Precision	468/546	86%

Table 5: La Hack 2 performance.

Parser

The parser we use has been developed over the past 6 months by Michael Collins, and is a continuation of the work on prepositional phrase attachment described in [5]. It was trained on 33000 sentences from the Wall Street Journal Treebank [8]. As yet no extensive performance tests have been made, but both recall and precision on labeled edges is over 80%. The parser was used to spot syntactic patterns which signaled coreference of noun phrases within sentences, such as appositive relations and predicate nominative constructions. The performance of this component is shown in table 6.

Given a maximal noun phrase, we find the head non-recursive noun phrase through a left-recursive descent of the parse tree. For example *Fred Bloggs, president of ACME, who was elected yesterday* would be reduced to *Fred Bloggs*. In addition, if either of the noun phrases involves conjunction, as in *president of General Motors and former CEO of Ford*, both minimal noun phrases, *president* and *former CEO* would be recovered.

We mark one noun phrase, called NP1, as being coreferent with a second noun phrase, NP2, because of an appositive relationship if NP1 is the head of a parent noun phrase, and NP2 is also a direct descendant of this parent noun phrase. For example, in the phrase *John Smith, president of ACME, a former worker at Eastern, John Smith* is coreferent with both *president* and *a former worker*. Note that the parser incorporates punctuation into the statistical model, so a comma between two noun phrases is seen as a strong indication of an appositive relationship.

The Wall Street Journal uses constructions similar to appositives to indicate relationships other than coreference. For example, such constructions are used with place names, such as *Frankfurt, Germany* or *Smith Barney, Harris Upham & Co. , New York* ; ages, such as *Al Bert, 49;* and dates, such as *March 31, 1989*. These constructions are a source of error in appositive recognition. In addition, the parser confuses some instances of conjunction with appositives. For this reason, semantic filtering is required to raise precision. We found that the following strategy worked remarkably well: given the two proposed minimal noun phrases, if the first one has a capitalized head, and the second head begins with a lower-case letter, accept the pair as coreferent. Note that this would deal correctly with all the above examples. A few additional cases were caught by allowing pairs where the first head word was on a list of honorifics, such as *president, chairman, journalist*, or *CEO*, and the second head was capitalized. This heuristic correctly handles cases such as *ACME's president, Bill Jones*. Also, a later processing stage removes indefinite cases from those proposed as appositives. While not appearing in the final output, these cases are used to aid in positing other types of coreference.

Definite cases of predicate nominative constructions are also markable. As a result, syntactic patterns of the type 'NP *is* NP' are also recognized, as are constructions involving the verbs *remain* or *become*, which function in a similar way to *be*. These could appear in sentential clauses or in relative clauses, such as *Fred Flintstone, who is Wilma's husband*. As is the case with appositives, indefinites are filtered from the final output, but are marked and used in later processing.

Several verbs function similarly to *become* and *remain*, but subcategorize for a prepositional phrase headed by *as*, with the object of this prepositional phrase being coreferent with the subject of the verb. A list of these verbs, including *serve, work, continue* and *resign*, was compiled and these patterns were used as well.

It was found that most verb phrases, regardless of the verb head, which take both a noun phrase, NP1, and a prepositional phrase headed by *as* with an object, NP2, imply coreference between NP1 and NP2. This was extended to include patterns of the form 'verb np1 (to be np2).' Some examples are shown below. Underlined entities are coreferent.

Mr. Casey succeeds M. James Barrett, 50, as president of Genetic Therapy

But the mainstream civil-rights leadership generally avoided the rhetoric of "law and order," regarding it as a code for keeping blacks back

We consider our Butthead to be an endearing, fun-loving guy," a spokesman says

In addition, patterns were implemented to identify phrases containing monetary figures in which alternate representations of the amount are present. Some such phrases are: *$53 , or 20 cents a share, 23 billion marks (15 billion dollars)* and *profits climbed to 11 million dollars.*

Recall	97/1627	6%
Precision	97/139	70%

Table 6: Syntactic Pattern Performance.

Parsing enables regular expressions to be written which apply to trees rather than surface text. These patterns are simpler and more intuitive than equivalent surface regular expressions. It is trivial to add new patterns to the system, since the parser has effectively abstracted away many of the complications of the surface text. While regular expressions could catch many of the phenomena we have described, they will become increasingly complex as they attempt to capture long range dependencies in the text and will also become increasingly inaccurate.

Bride of CogNIAC

Resolution of pronouns and lower-case anaphors was handled by a program called Bride of CogNIAC, which is an extension of CogNIAC, [1]. CogNIAC was designed to perform pronominal resolution in highly ambiguous contexts and is distinguished from other approaches to pronominal resolution in the following ways. First, it was designed to have high precision, rather than high recall. Second, it ranks the relative salience of an anaphor's candidate antecedents in a partial order rather than a total order. This means that two candidate antecedents can be equally salient. And, third, it requires that there be a unique antecedent for an anaphor. Uniqueness is achieved by eliminating competing antecedents using semantic information or by preferring some candidate antecedents over others. CogNIAC will not commit to a resolution if a unique referent cannot be found.

Bride of CogNIAC also handles lower-case definite descriptions using various knowledge sources to do semantic classification of noun phrases into categories such as person, male, female, place, thing, singular and plural. It also employs the pleonastic-it filter described above and a quoted speech component not present in CogNIAC. Bride of CogNIAC performs resolution on basal noun phrase detected and part-of-speech tagged text. It also relies on proper noun anaphora information provided by La Hack 2 and syntactic anaphora information posited by the parser. System performance prior to running Bride of CogNIAC, the last component which posits coreference, is shown in table 7.

Recall	564/1627	35%
Precision	564/648	82%

Table 7: La Hack 2 and Syntactic Pattern Performance.

Bride of CogNIAC attempts to determine whether fuzzy string matches such as *the unions* and *unions* indicate coreference. The combined performance of this component in conjunction with above components is shown in table 8. It equates markables which share a common head noun using various metrics of similarity. The biggest

difficulty is to prevent Bride of CogNIAC from marking too many things as coreferent. As a result, various heuristics are used to reduce the number of entities marked. For example, coreference is not posited if:

- The number of words in the antecedent noun phrase is less than the number of words in the anaphor.
- The words in either string are on a stop-word list.
- Possessive or prepositional modifier conflicts exist.

Recall	729/1627	45%
Precision	729/992	79%

Table 8: Performance with lower case string matching added.

The second and final task addressed by Bride of CogNIAC is the resolution of pronominals and words which behave like pronominals, such as *company*. Performance for this component alone is shown in table 9. Overall official results are shown in table 2. Overall unofficial results are shown in table 1.

Recall	245/1627	15%
Precision	245/423	58%

Table 9: Pronoun component performance.

We were disappointed by the performance of the pronoun resolution component. In examining the output briefly, the mistakes made were due to knowledge-base failures and bugs more than issues inherent to the pronoun resolution algorithm. This is clearly an aspect of the task where better knowledge representation would improve system performance.

CONCLUSION

We found the MUC-6 coreference task to be challenging and enjoyable for several reasons. First, most of us are accustomed to working alone and we enjoyed the opportunity to work as a team, especially since this fostered research contacts which might not have otherwise been made. Second, unlike typical research work, participation in MUC lasted a finite amount of time and there were clearly defined goals and success metrics. Third, the task exposed some of us to research areas with which we only had passing familiarity. We hope that MUC will continue to encourage participation from new sites by focusing on sub-tasks relevant to information extraction.

THANKS

We would especially like to thank Aravind Joshi and Mitch Marcus who allowed us to take time off from work directly related to our graduate studies to participate in the MUC-6 coreference task. We would also like to thank Mark Wasson for helpful comments on tokenization and for providing the inspiration for our data structure; Yael Ravin for giving us access to the Name Extraction Module and for documenting it so that we could quickly incorporate it into the coreference system; Peter Flynn for providing us with a large hand-built acronym dictionary which he maintains on a world wide web site in Iceland; and Christy Doran for analyzing some of the coreference data and providing us with many linguistic insights.

RESPONSE FILE FOR WALKTHROUGH ARTICLE

<DOC>
<DOCID> wsj94_026.0231 </DOCID>
<DOCNO> 940224-0133. </DOCNO>
<HL> marketing & media -- Advertising:
@ <COREF ID="3">John Dooner</COREF> will succeed <COREF ID="4">James</COREF>
@ at helm of <COREF ID="6">McCann-Erickson</COREF>
@ ----
@ by Kevin Goldman </HL>
<DD> 02/24/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B8 </SO>
<CO> IPG K </CO>
<IN> ADVERTISING (ADV), ALL ENTERTAINMENT & LEISURE (ENT),
FOOD PRODUCTS (FOD), FOOD PRODUCERS, EXCLUDING FISHING (OFP),
RECREATIONAL PRODUCTS & SERVICES (REC), TOYS (TMF) </IN>
<TXT>

One of the many differences between <COREF ID="11" TYPE="IDENT" REF="4">Robert L. James</COREF>, <COREF ID="12" TYPE="IDENT" REF="4">chairman</COREF> and <COREF ID="13" TYPE="IDENT" REF="4">chief executive officer</COREF> of <COREF ID="14" TYPE="IDENT" REF="6">McCann-Erickson</COREF>, and <COREF ID="15" TYPE="IDENT" REF="3">John J. Dooner Jr.</COREF>, <COREF ID="16">the agency</COREF>'s <COREF ID="17" TYPE="IDENT" REF="3">president</COREF> and <COREF ID="18" TYPE="IDENT" REF="3">chief operating officer</COREF>, is quite telling: <COREF ID="19" TYPE="IDENT" REF="4">Mr. James</COREF> enjoys sailboating, while <COREF ID="20" TYPE="IDENT" REF="3">Mr. Dooner</COREF> owns a powerboat.

Now, <COREF ID="22" TYPE="IDENT" REF="4">Mr. James</COREF> is preparing to sail into the sunset, and <COREF ID="24" TYPE="IDENT" REF="3">Mr. Dooner</COREF> is poised to rev up the engines to guide Interpublic Group's <COREF ID="27" TYPE="IDENT" REF="6">McCann-Erickson</COREF> into the 21st century. <COREF ID="29">Yesterday</COREF>, <COREF ID="30" TYPE="IDENT" REF="6">McCann</COREF> made official what had been widely anticipated: <COREF ID="32" TYPE="IDENT" REF="4">Mr. James</COREF>, 57 years old, is stepping down as chief executive officer on July 1 and will retire as chairman at the end of the year. <COREF ID="39">He</COREF> will be succeeded by <COREF ID="40" TYPE="IDENT" REF="3">Mr. Dooner</COREF>, 45.

It promises to be a smooth process, which is unusual given the volatile atmosphere of the <COREF ID="318">advertising</COREF> business. But <COREF ID="47" TYPE="IDENT" REF="3">Mr. Dooner</COREF> has <COREF ID="48">a big challenge</COREF> that will be <COREF ID="50" TYPE="IDENT" REF="3">his</COREF> <COREF ID="51" TYPE="IDENT" REF="48">top priority</COREF>. "<COREF ID="52" TYPE="IDENT" REF="39">I</COREF>'m going to focus on strengthening <COREF ID="53">the creative work</COREF>," <COREF ID="54" TYPE="IDENT" REF="39">he</COREF> says. "There is room to grow. <COREF ID="57">We</COREF> can make further improvements in terms of the perception of <COREF ID="61" TYPE="IDENT" REF="57">our</COREF> <COREF ID="62" TYPE="IDENT" REF="53">creative work</COREF>."

Even Alan Gottesman, an analyst with PaineWebber, who believes<COREF ID="67" TYPE="IDENT" REF="6">McCann</COREF> is filled with "vitality" and is in "great shape," says that from a creative standpoint, "You wouldn't pay to see <COREF ID="72" TYPE="IDENT" REF="6">their</COREF> reel" of <COREF ID="74">commercials</COREF>.

While <COREF ID="75" TYPE="IDENT" REF="6">McCann</COREF>'s world-wide billings rose 12% to $6.4 billion last year from $5.7 billion in 1992, <COREF ID="82" TYPE="IDENT" REF="16">the agency</COREF>

still is dogged by the loss of the key creative assignment for the prestigious Coca-Cola Classic account. "<COREF ID="86" TYPE="IDENT" REF="3">I</COREF> would be less than honest to say <COREF ID="87" TYPE="IDENT" REF="3">I</COREF>'m not disappointed not to be able to claim creative leadership for <COREF ID="89">Coke</COREF>,"<COREF ID="90" TYPE="IDENT" REF="3">Mr. Dooner</COREF> says.

<COREF ID="91" TYPE="IDENT" REF="6">McCann</COREF> still handles promotions a
nd media buying for <COREF ID="93" TYPE="IDENT" REF="89">Coke</COREF>. But the bragging rights to <COREF ID="95" TYPE="IDENT" REF="89">Coke</COREF>'s ubiquitous advertising belongs to <COREF ID="97">Creative Artists Agency</COREF>, <COREF ID="98" TYPE="IDENT" REF="97">the big Hollywood talent agency</COREF>. "<COREF ID="99" TYPE="IDENT" REF="57">We</COREF> are striving to have a strong renewed creative partnership with Coca-Cola," <COREF ID="102" TYPE="IDENT" REF="3">Mr. Dooner</COREF> says. However, odds of that happening are slim since <COREF ID="105">word</COREF> from <COREF ID="313" TYPE="IDENT" REF="89">Coke</COREF> headquarters in Atlanta is that <COREF ID="312" TYPE="IDENT" REF="97">CAA</COREF> and other ad agencies, such as Fallon McElligott, will continue to handle <COREF ID="314" TYPE="IDENT" REF="89">Coke</COREF> advertising.

<COREF ID="112" TYPE="IDENT" REF="3">Mr. Dooner</COREF>, who recently lost 60 pounds over <COREF ID="115" TYPE="IDENT" REF="39">three-and-a-half months</COREF>, says now that <COREF ID="116" TYPE="IDENT" REF="3">he</COREF> has "reinvented" himself, <COREF ID="118" TYPE="IDENT" REF="3">he</COREF> wants to do the same for <COREF ID="120" TYPE="IDENT" REF="16">the agency</COREF>. For <COREF ID="121" TYPE="IDENT" REF="3">Mr. Dooner</COREF>, it means maintaining <COREF ID="123" TYPE="IDENT" REF="3">his</COREF> running and exercise schedule, and for <COREF ID="125" TYPE="IDENT" REF="16">the agency</COREF>, <COREF ID="126" TYPE="IDENT" REF="16">it</COREF> means developing more global campaigns that nonetheless reflect local cultures. One <COREF ID="310" TYPE="IDENT" REF="6">McCann</COREF> account, "<COREF ID="131" TYPE="IDENT" REF="39">I</COREF> Can't Believe <COREF ID="132" TYPE="IDENT" REF="6">It</COREF>'s Not Butter," a butter substitute, is in 11 countries, for example.

<COREF ID="137" TYPE="IDENT" REF="6">McCann</COREF> has initiated a new so-called global collaborative system, composed of world-wide account directors paired with creative partners. In addition, Peter Kim was hired from WPP Group's J. Walter Thompson last September as vice chairman, chief strategy officer, world-wide.

<COREF ID="147" TYPE="IDENT" REF="3">Mr. Dooner</COREF> doesn't see a creative malaise permeating <COREF ID="149" TYPE="IDENT" REF="16">the agency</COREF>. <COREF ID="150" TYPE="IDENT" REF="3">He</COREF> points to several campaigns with pride, including <COREF ID="153">the Taster</COREF>'s Choice commercials that are like a running soap opera. "<COREF ID="157" TYPE="ID ENT" REF="153">It</COREF>'s a $19 million campaign with the recognition of a $200 million campaign," <COREF ID="161" TYPE="IDENT" REF="3">he</COREF> says of <COREF ID="162" TYPE="IDENT" REF="74">the commercials</COREF> that feature a couple that must hold a record for the length of <COREF ID="168">time</COREF> dating before kissing.

Even so, <COREF ID="170" TYPE="IDENT" REF="3">Mr. Dooner</COREF> is on the prowl for more creative talent and is interested in acquiring <COREF ID="173" TYPE="IDENT" REF="16">a hot agency</C OREF>. <COREF ID="174" TYPE="IDENT" REF="39">He</COREF> says <COREF ID="175" TYPE="IDENT" REF="39">he</COREF> would like to finalize <COREF ID="176" TYPE="IDENT" REF="29">an acquisition "yesterday</COREF>. <COREF ID="177" TYPE="IDENT" REF="39">I</COREF>'m not known for patience."

<COREF ID="179" TYPE="IDENT" REF="3">Mr. Dooner</COREF> met with <COREF ID="180">Martin Puris</COREF>, <COREF ID="181" TYPE="IDENT" REF="180">president</COREF> and <COREF ID="182"

TYPE="IDENT" REF="180">chief executive officer</COREF> of <COREF ID="183">Ammirati & Puris</COREF>, about <COREF ID="184" TYPE="IDENT" REF="6">McCann</COREF>'s acquiring the agency with billings of $400 million, but nothing has materialized. "There is no question," says <COREF ID="191" TYPE="IDENT" REF="3">Mr. Dooner</COREF>, "that <COREF ID="192" TYPE="IDENT" REF="57">we</COREF> are looking for quality acquisitions and <COREF ID="194" TYPE="IDENT" REF="183">Ammirati & Puris</COREF> is a quality operation. There are some people and entire agencies that <COREF ID="199" TYPE="IDENT" REF="3">I</COREF> would love to see be part of the <COREF ID="311" TYPE="IDENT" REF="6">McCann</COREF> family." <COREF ID="202" TYPE="IDENT" REF="3">Mr. Dooner</COREF> declines to identify possible acquisitions.

<COREF ID="204" TYPE="IDENT" REF="3">Mr. Dooner</COREF> is just gearing up for the headaches of running one of the largest world-wide agencies. (There are no immediate plans to replace <COREF ID="210" TYPE="IDENT" REF="3">Mr. Dooner</COREF> as <COREF ID="211" TYPE="IDENT" REF="3">president</COREF>; <COREF ID="212" TYPE="IDENT" REF="4">Mr. James</COREF> operated as chairman, chief executive officer and president for a period of <COREF ID="217" TYPE="IDENT" REF="168">time</COREF>.) <COREF ID="218" TYPE="IDENT" REF="4">Mr. James</COREF> is filled with thoughts of enjoying <COREF ID="220" TYPE="IDENT" REF="4">his</COREF> three hobbies: <COREF ID="222">sailing</COREF>, skiing and hunting.

Asked why <COREF ID="224" TYPE="IDENT" REF="4">he</COREF> would choose to voluntarily exit while <COREF ID="226" TYPE="IDENT" REF="4">he</COREF> still is so young, <COREF ID="227" TYPE="IDENT" REF="4">Mr. James</COREF> says it is <COREF ID="229" TYPE="IDENT" REF="168">time</COREF> to be a tad selfish about how <COREF ID="231" TYPE="IDENT" REF="4">he</COREF> spends <COREF ID="232" TYPE="IDENT" REF="4">his</COREF> days. <COREF ID="234" TYPE="IDENT" REF="4">Mr. James</COREF>, who has a reputation as <COREF ID="237">an extraordinarily tough taskmaster</COREF>, says that because <COREF ID="238" TYPE="IDENT" REF="4">he</COREF> "had <COREF ID="239" TYPE="IDENT" REF="168">a great time</COREF>" in <COREF ID="240" TYPE="IDENT" REF="318">advertising</COREF>," <COREF ID="241" TYPE="IDENT" REF="237">he</COREF> doesn't want to "talk about the disappointments." In fact, when <COREF ID="244">he</COREF> is asked <COREF ID="245" TYPE="IDENT" REF="244">his</COREF> opinion of the new batch of <COREF ID="315" TYPE="IDENT" REF="89">Coke</COREF> ads from <COREF ID="249" TYPE="IDENT" REF="97">CAA</COREF>, <COREF ID="250" TYPE="IDENT" REF="4">Mr. James</COREF> places <COREF ID="251" TYPE="IDENT" REF="4">his</COREF> hands over <COREF ID="253" TYPE="IDENT" REF="4">his</COREF> mouth. <COREF ID="255">He</COREF> shrugs. <COREF ID="256" TYPE="IDENT" REF="255">He</COREF> doesn't utter <COREF ID="257" TYPE="IDENT" REF="105">a word</COREF>. <COREF ID="258" TYPE="IDENT" REF="255">He</COREF> has, <COREF ID="259" TYPE="IDENT" REF="255">he</COREF> says, fond memories of working with <COREF ID="316" TYPE="IDENT" REF="89">Coke</COREF> executives. "<COREF ID="262" TYPE="IDENT" REF="89">Coke</COREF> has given <COREF ID="263" TYPE="IDENT" REF="57">us</COREF> great highs," says <COREF ID="265" TYPE="IDENT" REF="4">Mr. James</COREF>, sitting in <COREF ID="266" TYPE="IDENT" REF="255">his</COREF> plush office, filled with photographs of <COREF ID="269" TYPE="IDENT" REF="222">sailing</COREF> as well as huge models of, among other things, a Dutch tugboat.

<COREF ID="273">He</COREF> says <COREF ID="274" TYPE="IDENT" REF="273">he</COREF> feels a "great sense of accomplishment." In 36 countries, <COREF ID="278" TYPE="IDENT" REF="6">McCann</COREF> is ranked in the top three; in 75 countries, <COREF ID="281" TYPE="IDENT" REF="6">it</COREF> is in the top 10.

Soon, <COREF ID="283" TYPE="IDENT" REF="4">Mr. James</COREF> will be able to compete in as many sailing races as <COREF ID="285" TYPE="IDENT" REF="4">he</COREF> chooses. And concentrate on <COREF ID="286" TYPE="IDENT" REF="4">his</COREF> duties as rear commodore at the New York Yacht Club.

Maybe <COREF ID="290">he</COREF>'ll even leave something from <COREF ID="292" TYPE="IDENT" REF="290">his</COREF> office for <COREF ID="294" TYPE="IDENT" REF="3">Mr. Dooner</COREF>. Perhaps a framed page from <COREF ID="296">the New York Times</COREF>, dated Dec. 8, 1987, showing a year-end chart of the stock market crash earlier that year. <COREF ID="301" TYPE="IDENT" REF="4">Mr. James</COREF> says <COREF ID="302" TYPE="IDENT" REF="4">he</COREF> framed <COREF ID="303" TYPE="IDENT" REF="296">it</COREF> and kept <COREF ID="304" TYPE="IDENT" REF="296">it</COREF> by <COREF ID="305" TYPE="IDENT" REF="4">his</COREF> desk as a "personal reminder. <COREF ID="308" TYPE="IDENT" REF="296">It</COREF> can all be gone like that."
</p>
</TXT>
</DOC>

REFERENCES

[1] Baldwin, B. CogNIAC: A Discourse Processing Engine. University of Pennsylvania Department of Computer and Information Sciences Ph.D. Thesis, 1995.

[2] Brill, Eric. Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the Twelfth National Conference on AI*, (AAAI-94), Seattle, Washington, 1994.

[3] Byrd, R., Ravin, Y. and Prager, J. Lexical Assistance at the Information Retrieval User Interface. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, April 1995.

[4] Church, K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing*. February 1988.

[5] Collins, M. and Brooks, J. Prepositional Phrase Attachment through a Backed-off Model. In *Proceedings of the Third Workshop on Very Large Corpora*, June 1995.

[6] Karp, D., Schabes, Y., Zaidel, M. and Egedi, D.. A Freely Available Wide Coverage Morphological Analyzer for English. *Proceedings of the 15th International Conference on Computational Linguistics*, 1992.

[7] Lappin, S. and Leass, H. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, vol. 20, num. 4, pp. 538-539

[8] Marcus M., Santorini, B. and Marcinkiewicz, M., Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, vol. 19, num. 2, 1993.

[9] Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. Five Papers on WordNet. Cognitive Science Laboratory, Princeton University, No. 43, July 1990.

[10] Ramshaw, L. and Marcus, M. Text Chunking Using Transformation-Based Learning. In *Proceedings of the Third Workshop on Very Large Corpora*, June 1995.

[11] Ratnaparkhi, A. Forthcoming.