AN EVALUATION OF COREFERENCE RESOLUTION STRATEGIES FOR ACQUIRING ASSOCIATED INFORMATION

Lois C. Childs

Lockheed Martin Corporation

P.O. Box 8048
Philadelphia, PA 19101

lois@mds.lmco.com

(610) 354-5816

Category – Information Extraction

1. INTRODUCTION

As part of our TIPSTER research program [Contract Number 94–F133200–000], we have developed a variety of strategies to resolve coreferences within a free text document. Coreference is typically defined to mean the identification of noun phrases that refer to the same object. This paper investigates a more general view of coreference in which our automatic system identifies not only coreferential phrases, but also phrases which additionally describe an object. Coreference has been found to be an important component of many applications.

The following example illustrates a general view of coreference.

> *American Express, the large financial institution, also known as Amex, will open an office in Peking.*

In this example, we would like to associate the following information about American Express:

> its name is *American Express*;
> an alias for it is *Amex*;
> its location is *Peking, China*; and
> it can be described as *the large financial institution*.

In the work described in this paper, our goal was to evaluate the contributions of various techniques for associating an entity with three types of information:

1. Name Variations
2. Descriptive Phrases
3. Location Information

Data Set

The MUC6 Template Element task is typical of what our applications often require; it encapsulates in-formation about one entity within the Template Element. Since we have a way to evaluate our performance on this task via the MUC6 data, we used it to conduct our experiments. The corpus for the MUC6 Template Element task consists of approximately 200 documents for development (pre- and post-dry-run) and 100 documents for scoring. The scoring set had previously been held blind, but it has been released for the purposes of a thorough evaluation of our methods.

Scoring

Scores discussed in this paper measure performance of experimental system reconfigurations run on the 100 documents used for the final MUC6 evaluation. These scores were generated for inter-experiment comparison purposes, using the MUC6 scoring program, v1.3. Scores reported here are relevant only as relative measures within this paper and are not meant to represent official performance measures. Official MUC6 scores were generated using a later version of the scoring program. Furthermore, the scoring program results can vary depending on how the mapping between response and answer key is done. For example, if an automatic system has failed to make the link between a descriptor and a name, it may create two objects --- one for each. The scoring system must then decide which object to map to the answer key.

Object 1
NAME:	*American Express*
ALIAS:	*Amex*
TYPE:	COMPANY
LOCALE:	*Peking*
COUNTRY:	China

Object 2
DESCRIPTOR:	*the large financial institution*
TYPE:	COMPANY

Key

NAME:	*American Express*
ALIAS:	*Amex*
DESCRIPTOR:	*the large financial institution*
TYPE:	COMPANY
LOCALE:	*Peking*
COUNTRY:	China

The scoring program tries to optimize the scores during mapping but, if two objects would score equally, the scoring program chooses arbitrarily, thus, in effect, sacrificing a slot as a penalty for coreference failure. In the following example, the slot can be either NAME or DESCRIPTOR, depending on the mapping.

Object 1

NAME:	*American Express*
TYPE:	COMPANY

Object 2

DESCRIPTOR:	*the large financial institution*
TYPE:	COMPANY

Additionally, the answer key contains optional objects which are included in the scoring calculations only if they have been mapped to a response object. This sometimes causes a fluctuation in the number of possible correct answers, as reported by the scoring program. The scores, therefore, do not represent an absolute measure of performance.

Scores reported here use the following abbreviations:

POS	possible correct answers
ACT	actual answers produced
COR	correct answers
INC	incorrect answers
REC	recall (% of the correct answers found)
PRE	precision (% of answers found that are correct)

2. NAME VARIATIONS

Identifying variations of a person name or organization name is a basic form of coreference that underlies other strategies. Our process stores each newly recognized named entity, along with its computed variations and acronyms. The variations and acronyms are algorithmically generated without reference to the text. These are stored in a temporary lexicon so that variations of the name in the text can be recognized and linked to the original occurrence.

A careful examination of the name/alias results provides insight into the success of this technique.

Approximately two-thirds of the aliases were correctly identified. Of the one-third which were missed, besides an unfortunate system error which threw away four aliases which the system had found, five main groups of error were found. They can be categorized as follows:

1. Corporate Subsidiaries
2. Corporate Name Changes
3. Missing Name
4. Incomplete Name Variation
5. Unusual Firstname

Corporate Subsidiaries

There were approximately five missed aliases that involved corporations and their subsidiaries. In these cases, the aliases were assigned to the wrong entity. Usually, these were stories in which corporate officers were transferring from one part of a company to another. Confusion can quickly ensue when trying to link an alias with the correct entity in this case. (This is often true for the human reader, as well.) Find the three organizations in the following list of phrases:

EMI Records Group, a unit of London's Thorn EMI PLC

EMI Records Group North America

EMI Records Group

EMI

EMI Records

The three organizations are:

NAME:	*Thorn EMI PLC*
ALIAS:	*EMI*

NAME:	*EMI Records Group*
ALIAS:	*EMI Records*

NAME:	*EMI Records Group North America*

Of course, presentation of the names as a list is unfair to the reader because it eliminates all context cues. Rules which allow the automatic system to take greater advantage of context cues will be developed for such specialized areas.

Corporate Name Changes

Another five missed aliases were found in scenarios of changing corporate identity. By the rules of the Template Element task, the old name should become the alias of the new name. When these scenarios went un-

recognized by the system, the names were tagged as separate entities. The following is an example of a confusing name changing scenario which the automatic system missed.

> HEADLINE: Waste Management New Name
> Waste Management Inc. shareholders approved changing the name of this trash hauling, recycling and environmental services concern to WMX Technologies Inc.
> The company's North American solid-waste operations will retain the name Waste Management Inc.

The answer key for this scenario contains two organization entities.

> NAME: Waste Management Inc.
>
> and
>
> NAME: Waste Management Inc.
> or
> WMX Technologies Inc.
> ALIAS: Waste Management
> WMX Technologies Inc.
> or
> Waste Management
> Waste Management Inc.

Because there is some uncertainty within the text as to whether the change has already taken place, the second entity is given optional names covering both alternatives. This is difficult for an automatic extraction system to decipher.

Missing Name

Many aliases are found because they are variations of names which have been recognized by their form (i.e., they contain a corporate designator – Co.) or by their context (e.g., CEO of Atlas). Approximately ten missed aliases were due to the fact that the names themselves were not recognized. Improvement of name recognition is an on-going process as the system and its developers are exposed to more and more text.

Incomplete Name Variation

Name variations are generated algorithmically. There were only four aliases missed because they were not generated from the full name. Examination of the results has uncovered two new rules for making variations. These will be added to the set.

First, the abbreviation portion of the name should be included within an acronym, for example, ARCO as alias for Atlantic Richfield Company and RLA as alias for Rebuild L.A.

Second, a structural member like Chamber or Partnership can stand alone as a variation, for example, Chamber as alias for Chamber of Commerce and Partnership as alias for New York City Partnership.

It should be noted that our rule packages employ variable bindings to collect information during the pattern matching process. In the case of name variations, it would be helpful to tag the pattern's structural members that can stand alone as variants during the rule binding process. This can then guide the variation generator when that pattern has been matched.

Unusual Firstname

Seven PERSON aliases were missed because the system did not know the firstname, e.g. Clive, Vana, Rupert. The solution to this problem is not only to expand the system's knowledge of human firstnames, but also to widen the context which can trigger human name recognition. The system will be expanded to recognize as human those unknown words which are taking human roles, such as participating in family relationships.

Performance on the Name/Alias Task

Our system had the second highest score in organization alias identification in the MUC6 evaluation. (See the MUC6 Conference proceedings for official scores.)

ORGANIZATION ALIAS SCORE (V1.3)

POS	ACT	COR	INC	REC	PRE
170	153	110	2	65	72

Person alias scores were suppressed by 5 points of recall due to an error in the gender reference code. The following show the original scores and those after the error has been fixed.

PERSON ALIAS SCORE (V1.3) – ORIGINAL

POS	ACT	COR	INC	REC	PRE
170	157	146	1	86	93

PERSON ALIAS SCORE (V1.3) – ERROR FIXED

POS	ACT	COR	INC	REC	PRE
170	167	155	1	91	93

3. DESCRIPTIVE PHRASES

Associating an organization name with a descriptor requires resolving coreferences among names, noun phrases, and pronouns. Several techniques are involved here. Appositives, prenominals, and name-modified head nouns are directly associated with their respective

named entities during name recognition. After noun phrase recognition, those phrases which have not already been associated with a name are compared against known names in the text in order to find the correct referent.

Association by Context

During name recognition, entities are directly linked, via variable bindings within the patterns, with descriptive phrases that make up their context. This is a thrifty process because it allows the system to mine the very context which it has used to recognize the entity in the first place, thus allowing it to store linked information with the entity discovered. In this manner, the system is able to link descriptive phrases that are found in the following forms:

APPOSITIVE
MUCster Group, a New York consulting firm,

PRENOMINAL
the New York consulting firm, MUCster Group

NAME-MODIFIED HEAD NOUN
the MUCster Group consulting firm

Since the Template Element task described here restricted the descriptor slot to a single phrase, our system sought to choose the most reliable of all the phrases which had been linked to an entity. It did this by ranking the descriptors based on their syntactic role. The following is the ranking used for the MUC6 system:

1. appositive
2. predicate nominative
3. prenominal
4. name-modified head noun
5. longest descriptor (found by reference)

This ranking gives greater confidence to those descriptors associated by context, with the default choice, the longest descriptor, having been associated by reference.

70% of our system's name-linked descriptors were associated by context. This is not surprising in view of our ranked selection system. The following is a score of the original configuration, using the ranked selection system.

DESCRIPTOR SCORE (V1.3) – ORIGINAL CONFIGURATION

POS	ACT	COR	INC	REC	PRE
224	233	104	39	46	45

When the ranking is abandoned and the selection is based on the longest descriptor alone, 62% of the response descriptors are drawn from those associated by context. This change has a deleterious effect on the scores for the descriptor slot and confirms our hypothesis that the context-associated descriptors are more reliable.

DESCRIPTOR SCORE (V1.3) – LONGEST PREFERRED

POS	ACT	COR	INC	REC	PRE
223	233	87	53	39	37

A surprising result of this experiment is that the percentage of descriptors associated by context is still so high. This is believed to be due to their predominance within the set of noun phrases found by our system.

Association by Reference

Once an organization noun phrase has been recognized, the reference resolution module seeks to find its referent. This process involves several steps. First, the phrase is checked to make sure it hasn't already been associated by context. If not, a content filter for the phrase is run against a content filtered version of each known organization name; if there is a match, the link is made.

Content Filters:
"the jewelry chain"   =>(jewelry jewel chain)
"Smith Jewelers"   =>(smith jewelers jeweler jewel)

For example, if the organization noun phrase, "the jewelry chain" is identified, its content filter would be applied to the list of known company names. When it reaches "Smith Jewelers", it will compare the filter against a filtered version of the name. The best match is considered the referent. If there is a tie, file position is considered as a factor, the closest name being the most likely referent.

To assess the value of this filtering mechanism, the MUC6 evaluation corpus was processed without the filter. The following results show that the filter did help the system link the correct descriptors; without it, the system lost five points of recall and seven points of precision.

DESCRIPTOR SCORE (V1.3) – WITHOUT FILTER

POS	ACT	COR	INC	REC	PRE
222	235	90	48	41	38

For generic phrases like "the company" and for pronouns referring to people, reference is currently determined solely by file position and entity type. Plans have been formulated to increase the sophistication of this selection process, and to expand the system to handle coreference of pronouns to organizations.

Named vs. Un-named Organizations

Because of the possibility that a text may refer to an un-named organization by a noun phrase alone, it is necessary to recognize all definite and indefinite noun phrases that may refer to an organization. The following are examples of some un-named organizations:

the Clinton administration

federal appeals court

MUCster's coreference research group

a New York consultancy

its banking unit

an arm of the MUCster unit

Those phrases that have not already been associated with a named entity through context cues must then be associated by reference, if possible. For every definite noun phrase, if a reference can be found, it will be associated with that entity; otherwise, it will become an un-named entity. Every indefinite noun phrase that cannot be associated by context becomes an un-named entity.

During the filtering process, the system used an additional heuristic to decide whether to apply a content filter to a noun phrase, or to make it into an un-named entity. If a noun phrase is found to be especially rich in description, it is thought to be too specific to refer to a previous entity, and is made into an un-named entity. This heuristic turned out to be detrimental to performance; it suppressed the descriptor scores substantially. When the original configuration (i.e. ranked selection, favoring appositives) is run, without this heuristic, an increase of four recall and three precision points is achieved.

DESCRIPTOR SCORE (V1.3) – WITHOUT HEURISTIC

POS	ACT	COR	INC	REC	PRE
223	230	111	39	50	48

Context vs. Reference

The majority of descriptors reported were found through association by context, even when the "longest descriptor" selection method is used. This is partly due to the relative scarcity of unattached organizational noun phrases. Sixty-eight of the 224 possible descriptors were missed because the system did not recognize the noun phrase as describing an organization. When the key's descriptive phrases were added directly to the system's knowledge base, as a hard-coded rule package, to eliminate this variable, the following scores were produced.

DESCRIPTOR SCORE (V1.3) – ALL NOUN PHRASES ADDED

POS	ACT	COR	INC	REC	PRE
230	359	135	28	59	38

The responses scored were produced with the original system configuration which uses the ranked selection system. When the system reverts to preferring the longest descriptor, the following scores are achieved.

DESCRIPTOR SCORE (V1.3) – ALL NOUN PHRASES ADDED, LONGEST PREFERRED

POS	ACT	COR	INC	REC	PRE
230	366	132	31	57	36

The decline in scores adds further confirmation to our hypothesis that the context-associated descriptors are more reliable.

4. LOCATION INFORMATION

Finally, techniques for associating an organization name with location information are examined. This is an extension of traditional coreference, but a task we do in many applications. Biographical information about a person often falls into this category, e.g. address, telephone, or passport information. The intuition is that location information is found frequently within descriptive noun phrases and is extractable once that link has been established.

This approach was evaluated by examining the source of the answer key's locale fillers. It was found that 67% originated in appositives, prenominals, and post-modifiers, and 20% originated in other descriptive noun phrases.

APPOSITIVE
MUCster Group, a New York consulting firm,

PRENOMINAL
the New York consulting firm, MUCster Group

POST-MODIFIERS
MUCster Group (New York)
MUCster Group is based in New York

This may account for our system's superior performance in identifying locale/country information; our scores were the highest of the MUC6 participants. (See the MUC6 Conference proceedings for official scores.) We believe that this success is due to our method of collecting related information during name recognition.

LOCALE/COUNTRY SCORE (V1.3)

POS	ACT	COR	INC	REC	PRE
114	105	67	10	59	64
115	102	75	2	65	74

Breaking this down further, our system found 60% of those locale fillers which originated in prenominals, appositives, and post-modifiers, and 57% of the other 20%.

5. CONCLUSION

In the work described in this paper, our goal was to evaluate the contributions of various coreference resolution techniques for acquiring information associated with an entity. We looked at our system's performance in the MUC6 Template Element evaluation in three areas:

1. Name Variations
2. Descriptive Phrases
3. Location Information

Name Variations

Five areas were identified in which improvement to the name variation code is needed. Two areas will be improved by better modeling the events which may effect organizational names, e.g. the forming of subsidiaries and the changing of names. This can be extended to include other organizational events, such as corporate joint ventures. The third area, missing names, is an area of on-going improvement. Two new rules were identified to help the name variation algorithm. The last area of improvement, person names, can be improved on two fronts: 1) expanding the knowledge base of accepted first names, grouped by ethnic origin, and 2) better modeling frequent behaviors in which person names participate. The latter will be explored through automatic acquisition of person name context over a large corpus.

Despite the many areas for improvement that were identified, our system still had the second highest recall measure in organization alias, confirming the basic soundness of our approach.

Descriptive Phrases

Examination of our system's performance in associating descriptive phrases to a referent entity brought us to several conclusions regarding our system's techniques. First, our method of directly linking entities to the descriptive phrases that make up their context via variable bindings within patterns has been very successful. Second, the content filter does contribute to the effectiveness of our coreference resolution; its absence caused our scores to decline. It may be improved by expanding the filter to include semantic categories via a facility like WordNet, or through our internal conceptual hierarchy. Third, the heuristic that caused the system to discard phrases that it deemed too specific for resolution was extremely bad and costly to our performance. Fourth, our recognition of organizational noun phrases needs improvement. This may also benefit from a survey of typical contexts over a large corpus.

Location Information

Our system's success in identifying associated location information was due mainly to our method of collecting related information during name recognition, since 67% of the answer key's location information could be found within appositives, prenominals, and post-modifiers. As our methods of associating noun phrases by reference improves, our ability to associate location information may improve, as well.

Overall Performance

In summary, our system has incorporated many new techniques for associating coreferential information as part of our TIPSTER research program. This paper concludes that most of the techniques have been beneficial to our performance and suggests ways to further improvement.