Message Understanding Conference (MUC) Tests of Discourse Processing

Nancy A. Chinchor
Science Applications International Corporation
chinchor@gso.saic.com

Beth Sundheim
Naval Command, Control, and Ocean Surveillance Center, RDT&E Division (NRaD)
sundheim@nosc.mil

Abstract

Performance evaluations of NLP systems have been designed and conducted that require systems to extract certain prespecified information about events and entities. A single text may describe multiple events and entities, and the evaluation task requires the system to resolve references to produce the expected output.We describe an early attempt to use the results from an information extraction evaluation to provide insight notice relationship between the difficulty of discourse processing and performance on the information extraction task. We then discuss an upcoming noun phrase coreference evaluation that has been designed independently of any other evaluation task in order to establish a clear performance benchmark on a small set of discourse phenomena.

Background on the MUC Evaluations

Five Message Understanding Conferences have been held since 1987 (Sundheim and Chinchor 1993) and a sixth one is planned for 1995 (Grishman 1994). Each conference serves as a forum for reporting on a multisite evaluation of text understanding systems. Out of the experiences of the community of evaluators and evaluation participants has grown a basic paradigm of blackbox testing based on an information extraction task.

The basic paradigm consists of a task in which the systems under test are to process a designated set of texts to fill slots in a template database according to prespecified rules. The domain of the test and the prespecified rules are developed with the interests of the research community (technical challenge), the evaluators (evaluability), and the potential customers (utility) in mind. For example, the domain of MUC-3 and MUC-4 (Chinchor, Hirschman, and Lewis 1993) was terrorist activity in nine Latin American countries. The systems had to analyze news articles and determine whether a reported event was a terrorist event and, if so, who had done what to whom. The systems then put this information into a template containing slots such as event type, perpetrator, target, and effect. A study of discourse-related aspects of the MUC-4 information extraction task (Hirschman 1992) is summarized in this paper.

The MUC-6 evaluation, which is scheduled to be conducted in the fall of 1995, will include a modified version of an information extraction task, and it will also include two text-tagging tasks. One of the text-tagging tasks is to identify some types of coreference relations. The design of this task is described in a later section of this paper.

Testing Event Tracking in an Information Extraction Context

Representatives of eight MUC-3 sites authored a joint paper on discourse processing in the MUC-3 systems (Iwanska et al 1991). The paper describes the capabilities of the MUC-3 systems in the following three areas:

1. Identifying portions of text that describe different domain events (recognizing single event vs. multiple events)
2. Resolving references:
   a. pronoun references
   b. proper name references
   c. definite references
3. Discourse representation

The group concluded that the tasks of recognizing a single event and distinguishing multiple events were the most important aspects of discourse-related processing for the information extraction task. While most systems did do reference resolution, they had various ways of doing it. Most systems did not produce an explicit discourse representation.

All but one author believed that handling discourse-related phenomena was the area which would yield the most improvement in performance. Extensions to the following processes and data were believed to be means of recognizing a single event and distinguishing multiple events: reference resolution (particularly, distributed definite anaphora and vague references), temporal and spatial reasoning, ambiguity resolution, semantic criteria for merging events, and general world knowledge.

To explore further the effects of discourse on perfor-

mance of an information extraction task, Hirschman (Hirschman 1992) carried out an adjunct test during the MUC-4 evaluation. The test was based on the information distribution in the input texts and the output templates. Two hypotheses were posed:

1. *The Source Complexity Hypothesis*
   The more complex the distribution of the source information for filling a given slot or template (the more sentences, and the more widely separated the sentences), the more difficult it will be to process the message correctly.
2. *The Output Complexity Hypothesis*
   The more complex the output (in terms of number of templates), the harder it will be to process the message correctly.

The focus of the test was the event merger problem: deciding whether two clauses describe the same event or distinct events. Two kinds of errors were identified:

1. *Lazy Merger*
   Two clauses describe a single event and should be merged at the template level, but the system fails to merge them.
2. *Greedy Merger*
   Two clauses describe two different events and should not have been merged.

Four subsets of texts from the MUC-4 test set were created according to which of these problems would or would not arise. The subset for which neither of these problems would arise -- the one whose messages generate one template whose fill derives from a single sentence -- did not turn out to be the easiest. It included many passing references to relevant events, which systems had trouble detecting. However, the subset where both problems could arise was indeed harder than the others. There were indications that both lazy merger and greedy merger were real problems for discourse processing. Another result was the observation that the performance across systems on this test was relatively consistent with respect to the message subsets, suggesting that some texts may simply be harder than others.

The difficulties of doing blackbox testing to isolate one stage of processing for testing were apparent. The small amount of data available after all other factors are filtered out makes it difficult to rely on the conclusions of this test. However, despite the fact that the test design obscured some of the discourse issues of interest, it provided some unexpected and interesting insights into what may cause some messages to be more difficult to analyze than others.

Testing Coreferential NP Identification as a Text-Tagging Task

The MUC-6 evaluation allows participation by sites in one or more of several component evaluations. One of the component evaluation tasks is to identify coreference relations. The initial concept for this task was developed by Jerry Hobbs (SRI International); the concept is currently being actively discussed and refined under the leadership of Ralph Grishman (NYU) by the MUC-6 planning committee including representatives from several sites (BBN Systems and Technologies, Brandeis University, Durham University, Martin Marietta Management and Data Systems, The MITRE Corporation, New Mexico State University, University of Pennsylvania, PRC Inc., Sheffield University, Southern Methodist University, SRA Corporation, SRI International, Unisys Corporation).

The development of this test started with a method of notating as many cases of noun phrase coreference as could be found in the texts. To that end, Hobbs proposed the following task:

> For every name, pronoun, or noun phrase that is coreferential with or inferable from something earlier in the text, specify the referring expression X, the antecedent Y, and the relation between them.Virtually any relation is possible, but several relations are very common. The latter should be labeled as such. These relations are
>
> (Ident X Y): X is identical to Y
>
> (Sub X Y): X is a subset or element of Y
>
> (Sub Y X): Y is a subset or element of X
>
> (l-subj Y X): Y is the logical subject of a designated nominalization X
>
> (l-obj Y X): Y is the logical object of a designated nominalization X
>
> Otherwise, the relation is
>
> (Rel X Y)

In just a few example sentences from Wall Street Journal articles, Grishman found the following types of coreference to annotate:

pronoun coreference
definite NP coreference
name coreference
apposition
NP coreference
implicit argument of nominalization
explicit argument of nominalization
control

Based on feedback from committee members, who had annotated a small number of articles by hand, it was decided that the MUC-6 effort should be limited to a small subset of types and relations that could be annotated consistently and explained clearly to evaluation participants. It was also decided that the task should be a text annotation task using SGML compliant markup. The answer key will be produced using an annotation tool developed by SRA Corporation (Aone and Bennett 1994). The tool facilitate uniformity, cuts down on human error, and decreases the time required to do the annotation. Also, a scoring tool developed by SAIC (Chinchor 1995) will be used to automatically score results and to test interannotator consistency.

Summary Description of the MUC-6 Coreference Task Annotation (Version 1.1)

The annotation for coreference is SGML tagging within the text stream (based on Aone and McKee 1993). The notation allows the expression of a link between an explicitly marked anaphor and antecedent.

(1) <COREF ID="100">Lawson Mardon Group Ltd.</COREF> said <COREF ID="101" TYPE="IDENT" REF="100">it</COREF>...

In the above example, the pronoun "it" is tagged as referring to the same entity as the phrase, "Lawson Mardon Group Ltd." The purpose of the TYPE attribute is to indicate the relationship between the anaphor and the antecedent. Only some of the possible relationships that can hold are captured by the notation. The TYPE attribute can have one of five values:

IDENT for "identical,"

PT-WH for "part/whole,"

WH-PT for "whole/part,"

SUB-SUP for "subset/superset," or

SUP-SUB for "superset/subset."

The PT-WH and WH-PT types and the SUB-SUP and SUP-SUB types indicate not only what the basic type of coreference relation is but also which role the current string plays in the coreference relation. An example of WH-PT follows:

(2) <COREF ID="100">Toledo</COREF>

... <COREF ID="101" TYPE="WH-PT" REF="100">the country</COREF>

When a referring expression has two or more antecedents, as in (3) below, the REF of the referring expression is multivalued, i.e., it contains the indices of each of its antecedents, and the REF values are listed, as shown.

(3) <COREF ID="100">Lawson Mardon Group Ltd.</COREF>

... <COREF ID="101">MB Group PLC</COREF>

... <COREF ID="102" TYPE="SUP-SUB" REF="100 101">the two companies</COREF>

In (4) an example of an optional tag and indication of minimum strings is given. Marking of coreference in the predicate nominative is discussed later.

(4) <COREF ID="102" MIN="decline">The August decline</COREF> followed a drop of $597 million in July, and is <COREF ID="103" STAT="OPT" TYPE="IDENT" REF="102" MIN="decrease">the fourth consecutive monthly decrease</COREF>.

It should be noted here that the annotation results in coreference "chains" by linking a "mention" (our meaning of "antecedent") and a "subsequent mention" (our meaning of "anaphor") and so on (later "anaphors").

Discussion of Scoring of MUC-6 Coreference Task

To define a scoring scheme for the coreference markup, the questions of what metrics to use and how to formulate them need to be answered. Special formulations of the "recall" and "precision" metrics that have been used for information extraction tasks will be applied to the coreference task. Recall measures the percentage of correct information produced by the system under test. Precision measures the percentage of information produced by the system that is correct. The metrics are in tension with each other, as any attempt by the system to generate *additional* information runs the risk of generating incorrect information rather than correct information. Thus, there is the tendency for recall to improve at the expense of precision or for precision to improve at the expense of recall.

Formulation of the metrics can take advantage of the information captured in the markup concerning the linkages among strings and the coreference types associated with the links. The coreference markup was designed to facilitate scoring. Links between anaphors and antecedents are captured in the markup, but scoring should not be based on the number of links that match between the answer key ("key") and the system-generated output ("response"), because the key and the response may have different ways of annotating equivalent coreferential relationships. For example, the key could mark A as identical to B and B as identical to C, while the response could mark A as identical to B and A as identical to C. Since the identity relationship is transitive, the same equivalence classes result from both ways of marking up the text. If computation of scores were based on the number of matching links, the response would not be

fully correct even though it is equivalent to the key.

Instead, for the identity relationship, scoring should be based on a comparison of the equivalence classes defined by the links in the key and the response, as described in a technical report by The MITRE Corporation (Burger et al 1994). MITRE's model-theoretic approach to scoring identity coreference assigns a recall score for a text based on the minimal number of changes required to transform the equivalence classes of the response into those of the key. It assigns that score using a computationally cheap algorithm that is a very simple counting scheme. It is based on taking each equivalence set defined by the key and partitioning it to find "islands" by intersecting it with the equivalence sets defined by the response. Using a simple formula requiring only the calculation of the cardinalities of equivalence sets in the key and the partitions, the "islands" can be connected in a minimal way and recall can be calculated. Precision can be calculated likewise by switching the "figure" and the "ground" and assigning a precision score based on the minimal number of changes required to transform the equivalence classes of the *key* into those of the *response*.

The MITRE solution is elegant, but the problem of scoring linkages other than IDENT still requires a more costly approach, which has been worked out with the assistance of Jerry Hobbs (SRI International). It involves counting resolved, subsequent NPs in the coreference chain that are determined to be "relationally equivalent." Recall is the percentage of resolved, subsequent NPs in the key that are determined to be relationally equivalent to x in the key, determined to be relationally equivalent to y in the response, and for which there exists a z such that x is determined to be relationally equivalent to z in the key and y is relationally equivalent to z in the response. Likewise, precision is defined but with key and response reversed. This algorithm extends the notion of equivalence sets to other relations, but requires more computing power than the model-theoretic approach even if IDENT is the only relation involved.

Training Data and Interannotator Agreement

A set of 318 Wall Street Journal articles on the ACL/DCIdisk obtained from the Linguistic Data Consortium by each site represented on the planning committee was annotated as data for evaluation participants to use during system development and as data for discussion of the issues raised by the application of the task guidelines. The articles were parceled out among the 16 sites with overlaps between assignments. Interannotator scoring was done to determine the amount of agreement on overlapping assignments that was achieved on the identity relation for the task as defined in Version 1.1. It was found that recall overall for the answer keys was 46% and precision was 54%. Slightly higher scores were achieved when "optional" and "minimum" markings were not checked; overall recall was 59% and precision 71%. However, this is not good enough for an answer key and indicates that the task is overly ambitious and/or underspecified. We expect at least 80% agreement among humans for the test to be a fair one for machines (Will 1993, Sundheim 1991, Sundheim 1992). Based on experience gained during the collaborative annotation effort and on the results of the interannotator agreement test, the committee has engaged in discussion centered around ways to limit the cases of coreferent NPs covered by the task and to better define the coreference relations.

Issues in the Version 1.1 Definition of the Coreference Task

Relations

There are two main kinds of issues that arise concerning the coreference relations to be tagged. First, the relations need to have clear definitions and a number of examples. Second, we need to use our clearer understanding of the current guidelines to help us to decide how to simplify or reduce the task. Proposals considered include conflating part-whole and subset-superset, eliminating these relations and restricting tagging to identity, or limiting tagging to certain cases of the relations. The current consensus of opinion is that subset-superset and part-whole should not be conflated because they are clearly two separate and definable relationships. The preference is also that these relations not be eliminated if at all possible because they are important for applications such as information extraction. For example, it would be important for extraction applications to identify a part-whole relation between a subsidiary organization and the parent organization. It is agreed that more work by the committee should be done to better define the relations and to produce examples to guide the annotators in another round of annotations to see if the consistency of the markings can be improved. As for limiting cases, the major proposals center around semantic classes, grammatical classes, and syntactic structures.

Semantic Classes

The best suggestion made for simplification of the task may be to restrict the tagging to people, places, and organizations because the challenges in NP coreference tagging remain, but the amount of work in preparing the answer keys is reduced. Furthermore, the definition of the classes is already quite clear and well supported by examples as a result of work carried out by the committee on another MUC-6 task, called "Named Entity" (Sundheim 1994). At this point in the discussion, the committee has considered the suggestion briefly but is focusing on other possible

ways to simplify and improve the evaluation design.

Grammatical Classes and Syntactic Structures

It has been proposed that only non-sentential NPs o be marked as anaphors and antecedents. This restriction on grammatical classes allows us to avoid difficult issues raised by many coreference phenomena, as in the following example:

(5) Program trading is "a racket," complains Edward Egnuss, a White Plains, N.Y., investor and electronics sales executive, "and it's not to the benefit of the small investor, that's for sure."

Though "that" is related to "it's not to the benefit of the small investor", the latter is not an NP, so the link is not annotated.

The only other syntactic restrictions on which the committee has reached consensus is to not tag relative pronouns and their heads, because the identity is automatic given the syntax, at least in most cases. However, relations encoded in possessives, partitives, appositives and NP predicate adjuncts should be annotated because of the adverse consequences of *not* annotating them, namely the creation of gaps in the coreference chains and the impact of those gaps on other coreference relations. In general, then, syntactic restrictions will not produce much meaningful simplification of the task. In fact, version 1.1 of the task definition disallowed tagging of coreference in a wide range of syntactically defined contexts and these guidelines will be revised to allow a broader range of phenomena to fall within the scope of the task. We have found the current guidelines hard for people to follow, in addition to finding that adherence to them results in gaps in the coreference chains.

Heads versus Full NPs

The proposal to tag head nouns as opposed to full NPs is currently being discussed. Those in favor of it feel that tagging head nouns simplifies the task for humans and neither oversimplifies nor overcomplicates the task for machines. Those opposed feel that it complicates the task and that it will not lead to greater interannotator consistency. Problems arise in agreeing on a definition of the head for various types of NP structures and in agreeing on whether systems can more reliably identify the head than the full NP. There seems to be evidence that, even if we decide to tag only heads, we will still need to indicate a longer, alternative string in the answer key for some ambiguous cases such as the extent of a name. An additional argument against tagging only the head is that information from the full NP is required to resolve coreference. At the current moment, the committee has not reached consensus regarding this issue.

Summary

The MUC-4 evaluation of event tracking in the context of an information extraction task provided limited insight into discourse issues that had been identified by MUC-3 evaluation participants as crucially affecting system performance on an information extraction task. The MUC-6 coreference task is being defined independently of an information extraction task and should therefore provide significantly more insight into a more limited set of coreference phenomena involving noun phrases. The exercise of defining the task and reaching consistency in annotating texts will produce research results in itself. The task design should provide a baseline for more ambitious future efforts, and the evaluation results should establish a performance benchmark that will be useful to the research community as well as to potential technology consumers.

References

Aone, C. and S.W. Bennett (1994)"Discourse Tagging Tool and Discourse-tagged Multilingual Corpora" in *Proceedings of International Workshop on Sharable Natural Language Resources*. Nara, Japan.

Aone, C. and D. McKee (1993) "Language-Independent Anaphora Resolution System for Understanding Multilingual Texts" in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, OH. p. 162.

Burger, J., M. Vilain, J. Aberdeen, D. Connolly, and L. Hirschman (1994) "A Model-Theoretic Coreference Scoring Scheme" Technical Report. The MITRE Corporation.

Chinchor, N., L. Hirschman, and D. Lewis (1993) "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)" *Computational Linguistics* 19(3). pp. 409 - 449.

Chinchor, N. (1995) *MUC-6 Scoring System User's Manual*. Technical Report. SAIC. San Diego, CA.

Grishman, R. (1994) "Wither Written Language Evaluation?" in *Proceedings: ARPA Workshop on Human Language Technology*. pp. 116 - 121.

Hirschman, L. (1992) "An Adjunct Test for Discourse Processing in MUC-4" in *Proceedings: Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufman Publishers, Inc. San Francisco, CA. pp. 67 - 77.

Iwanska, L., D. Appelt, D. Ayuso, K. Dahlgren, B. Glover

Stalls, R. Grishman, G. Krupka, C. Montgomery, and E. Riloff (1991) "Computational Aspects of Discourse in the Context of MUC-3" in *Proceedings: Third Message Understanding Conference (MUC-3).* Morgan Kaufman Publishers, Inc. San Francisco, CA. pp. 256 - 282.

Sundheim, B. (1991) "Overview of the Third Message Understanding Evaluation and Conference" in *Proceedings: Third Message Understanding Conference (MUC-3).* Morgan Kaufman Publishers, Inc. San Francisco, CA. p. 12.

Sundheim, B. (1992) "Overview of the Fourth Message Understanding Evaluation and Conference" in *Proceedings: Fourth Message Understanding Conference (MUC-4).* Morgan Kaufman Publishers, Inc. San Francisco, CA. p. 18.

Sundheim, B. and N. Chinchor (1993) "Survey of the Message Understanding Conferences" in *Proceedings: Human Language Technology.* Morgan Kaufman Publishers, Inc. San Francisco, CA. pp. 56 - 60.

Sundheim, B. (1994) *Named Entity Task Definition.* Technical Report. Naval Command, Control, and Ocean Surveillance Center, RDT&E Division (NRaD). San Diego, CA.

Will, C. A. (1993) "Comparing Human and Machine Performance for Natural Language Information Extraction: Results for English Microelectronics from the MUC-5 Evaluation" in *Proceedings: Fifth Message Understanding Conference (MUC-5).* Morgan Kaufman Publishers, Inc. San Francisco, CA. pp. 53 - 67.