Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 1-26
©Comptational Linguistics Society of ROC

Analyzing the Performance of Message Understanding Systems

Amit Bagga[*] and Alan W. Biermann[*]

Abstract

In this paper we describe a method of classifying facts (information) into categories or levels; where each level signifies a different degree of difficulty of extracting the fact from a piece of text containing it. Based on this classification mechanism, we propose a method of evaluating a domain by assigning to it a "domain number" based on the levels of a set of *standard* facts present in the articles of that domain. We then use the classification mechanism to analyze the performances of three MUC systems (BBN, NYU, and SRI) based on their ability to extract a set of *standard* facts (at different levels) from two different MUC domains. This analysis is then extended to analyze the role of coreferencing in the performance of message understanding systems.

The evaluation of a domain based on the "domain number" assigned to it is a big step up from methods used earlier (which used vocabulary size, average sentence length, the number of sentences per document, etc.). Moreover, the use of the classification mechanism as a tool to analyze the performance of message understanding systems provides a *deeper* insight into these systems than the one provided by obtaining the precision and recall statistics of each system.

Keywords: Information Extraction, Domain Complexity, Analysis of Systems, Message Understanding Conferences

1. Introduction

The Message Understanding Conferences (MUCs) have been held with the goal of qualitatively evaluating message understanding systems. The six MUCs held thus far have been quite successful at providing such an evaluation. Since MUC-3, the systems have been evaluated on three different domains, and the task has been expanded from

[*]Department of Computer Science, Duke University, Durham, NC 27708-0129, USA.
E-mail: {amit, awb}@cs.duke.edu

simply filling templates, in MUC-3 [MUC-3 1991], to including named entity recognition (NE) and coreferencing (CO), in MUC-6 [MUC-6 1995], as well. For MUC-6, the precision statistics of the participating systems varied from 34% to 73% and the recall statistics varied from 32% to 58% on the scenario template (ST) task.

But while the MUCs have shown the differences in the performance of the systems for a particular task (in a particular domain), little or no work has been done in trying to explain the differences in the performance of the systems. In addition, very little work has been done in analyzing the difficulty of understanding a text in a particular domain; both, independently, as well as in comparison to understanding a text in some other domain.

The organizers of MUC-5 attempted to compare the difficulty of the EJV (English Joint Ventures) task in MUC-5 to the terrorist task of MUC-3 and MUC-4. The criteria used for comparing these two tasks included the vocabulary size, the average sentence length, the average number of sentences per text, the number of texts, etc. [Sundheim 1993]. The organizers of MUC-6 did not attempt to compare the difficulty of the MUC-6 task to the previous MUC tasks saying that "the problem of coming up with a reasonable, objective way of measuring relative task difficulty has not been adequately addressed" [Sundheim 1995].

In this paper we describe a method of classifying facts (information) into categories or levels; where each level signifies a different degree of difficulty of extracting the fact from a piece of text containing it. Based on this classification mechanism, we propose a method of evaluating a domain by assigning to it a "domain number" based on the levels of a set of *standard* facts present in the articles of that domain. We then use the classification mechanism to analyze the performances of three MUC systems (BBN, NYU, and SRI) based on their ability to extract a set of *standard* facts (at different levels) from two different MUC domains. This analysis is then extended to analyze the role of coreferencing in the performance of message understanding systems.

2. Definitions

Semantic Network:

A *semantic network* consists of a collection of nodes interconnected by an accompanying set of arcs. Each node denotes an object and each arc represents a binary relation between the objects [Hendrix 1979].

A Partial Semantic Network:

A *partial semantic network* is a collection of nodes interconnected by an accompanying set of arcs where the collection of nodes is a subset of a collection of nodes forming a semantic network, and the accompanying set of arcs is a subset of the set of arcs accompanying the set of nodes which form the semantic network.

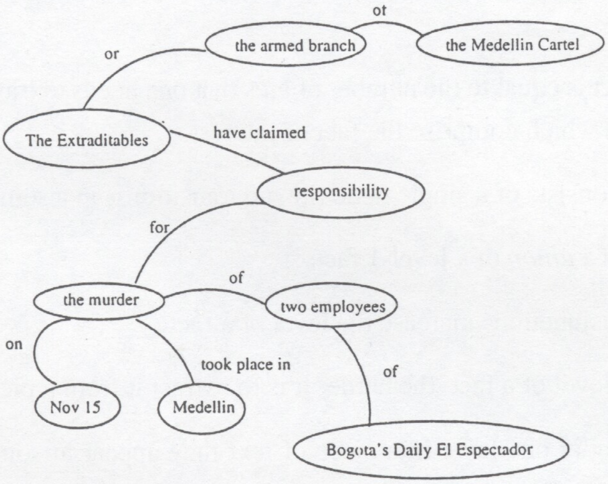

Figure 1 A Sample Semantic Network

Figure 1 shows a sample semantic network for the following piece of text:

> "The Extraditables," or the Armed Branch of the Medellin Cartel have claimed responsibility for the murder of two employees of Bogota's daily El Espectador on Nov 15. The murders took place in Medellin.

3. The Level of A Fact

The level of a fact, *F*, in a piece of text is defined by the following algorithm:

(1) Build a semantic network, *S*, for the piece of text.

(2) Suppose the fact, *F*, consists of several nodes $\{x_1, x_2, ..., x_n\}$. Let *s* be the partial semantic network consisting of the set of nodes $\{x_1, x_2, ..., x_n\}$ interconnected by the set of arcs $\{t_1, t_2, ..., t_k\}$.

We define the *level* of the fact, *F*, *with respect to* the semantic network, *S* to be equal to *k*, the number of arcs linking the nodes which comprise the fact *F*.

3.1 Observations

Given the definition of the level of a fact, the following observations can be made:

- The level of a fact is related to the concept of "semantic vicinity" defined by Schubert et al. [Schubert 1979]. The *semantic vicinity* of a node in a semantic net consists of the nodes and the arcs reachable from that node by traversing a small number of arcs. The fundamental assumption used here is that "the knowledge required to perform an intellectual task generally lies in the semantic vicinity of the concepts involved in the task" [Schubert 1979].

The level of a fact is equal to the number of arcs that one needs to traverse to reach all the concepts (nodes) which comprise the fact of interest.

- A level-0 fact consists of a single node (i.e. no transitions) in a semantic network.

- A level-*k* fact is a *union* of *k* level-1 facts.

- Conjunctions/disjunctions increase the level of a fact.

- The higher the level of a fact, the harder it is to extract it from a piece of text.

- A fact appearing at one level in a piece of text may appear at some other level in the same piece of text.

- The level of a fact in a piece of text depends on the granularity of the semantic network constructed for that piece of text. Therefore, the level of a fact with respect to a semantic network built at the word level (i.e. words represent objects and the relationships between the objects) will be greater than the level of a fact with respect to a semantic network built at the phrase level (i.e. noun groups represent objects while verb groups and preposition groups represent the relationships between the objects).

3.2 Examples

Let *S* be the semantic network shown in Figure 1. *S* has been built at the phrase level.

- The city mentioned, in *S*, is an example of a level-0 fact because the "city" fact consists only of one node "Medellin."

- The type of attack, in *S*, is an example of a level-1 fact.

We define the *type of attack* in the semantic network to be an attack designator such as "murder," "bombing," or "assassination" with one modifier giving the victim, perpetrator, date, location, or other information.

In this case the type of attack fact is composed of the "the murder" and the "two employees" nodes and their connector. This makes the type of attack a level-1 fact.

The type of attack could appear as a level-0 fact as in "the Medellin bombing" (assuming that the semantic network is built at the phrase level) because in this case both the attack designator (bombing) and the modifier (Medellin) occur in the same node. The type of attack fact occurs as a level-2 fact in the following sentence (once again assuming that the semantic network is built at the phrase level): "10 people were killed in the offensive which included several bombings." In this case there is no direct connector between the attack designator (several bombings) and its modifier (10 people). They are connected by the intermediatory "the offensive" node; thereby making the type of attack a level-2 fact. The type of attack can also appear at higher levels.

•In *S*, the date of the murder of the two employees is an example of a level-2 fact. This is because the attack designator (the murder) along with its modifier (two employees) account for one level and the arc to "Nov 15" accounts for the second level.

The date of the attack, in this case, is not a level-1 fact (because of the two nodes "the murder" and "Nov 15") because the phrase "the murder on Nov 15" does not tell one that an attack actually took place. The article could have been talking about a seminar on murders that took place on Nov 15 and not about the murder of two employees which took place then.

•In *S*, the location of the murder of the two employees is an example of a level-2 fact. The exact same argument as the date of the murder of the two employees applies here.

•The complete information, in *S*, about the victims is an example of a level-2 fact because to know that two employees of Bogota's Daily El Espectador were victims, one has to know that they were murdered. The attack designator (the murder) with its modifier (two employees) accounts for one level, while the connector between "two employees" and "Bogota's Daily El Espectador" accounts for the other.

•Similarly, the complete information, in *S*, about the perpetrators of the murder of the two employees is an example of a level-5 fact. The breakup of the 5 levels is as follows: the fact that two employees were murdered accounts for one level; the fact that "The Extraditables" have claimed responsibility for the murders accounts for two additional levels; and the fact that the Extraditables are the "armed branch of the Medellin Cartel" account for the remaining two levels.

4. Justification of the Methodology

The level of a fact quantifies the "spread" in the information that makes up the fact. Therefore, the higher the level of a fact, the greater is the "spread" in the information that makes up the fact. This means that more processing has to be done to identify and link all the individual pieces of information that make up the fact. In fact, an exploratory study done by Beth Sundheim during MUC-3 showed "a degradation in correctness of message processing as the information distribution in the message became more complex, that is, as slot fills were drawn from larger portions of the message and required more discourse processing to extract the information and reassemble it correctly in the required template(s)" [Hirschman 92].

An argument can be made that there are other factors, apart from the spread of information, which influence the difficulty of extracting a fact from text. Some of these factors include the amount of training done on an information extraction system, the quality of training, and the frequency of occurrence of the patterns that a system has been trained on. While these factors do influence the performance of an information extraction system and they do give some indication as to how difficult it was for a particular system to extract the fact, they do not give a system independent way of determining the complexity of extracting the fact.

In [Hirschman 92], Lynette Hirschman proposed the following hypothesis: there are facts that are simply harder to extract, across all systems. Based on our definition of the level of a fact, we analyzed the performances of three different information extraction systems on the MUC-4 terrorist reports domain and the MUC-6 management changes domain. Our analysis shows that all the three systems consistently did much worse on higher level facts in both the domains. In addition to confirming Hirschman's hypothesis, the analysis also shows that higher level facts are indeed harder to extract. Full details of the analysis are given later in this paper.

5. Building the Semantic Networks

As mentioned earlier, the level of a fact for a piece of text depends on the semantic network constructed for the text. Since there is no unique semantic network corresponding to a piece of text, care has to be taken so that the semantic networks are built consistently.

For the set of experiments described in the rest of the paper we used the following algorithm to build the semantic networks:

(1) Every article was broken up into a non-overlapping sequence of noun groups (NGs),

verb groups (VGs), and preposition groups (PGs). The rules employed to identify the NGs, VGs, and PGs were almost the same as the ones employed by SRI's FASTUS system.[1]

Full parsing of English sentences is AI-complete. However, certain syntactic constructs can be reliably identified. One such construct is a noun group, that is, a noun group is a noun phrase up to the head noun. Another is the verb group, that is, the verb together with its auxiliaries and embedded adverbs. A preposition group consists of single prepositions.

(2) The nodes of the semantic network consisted of the NGs while the transitions between the nodes consisted of the VGs and the PGs.

(3) Identification of coreferent nodes and prepositional phrase attachments were done manually.

Obviously, if one were to employ a different algorithm for building the semantic networks, one would get different numbers for the level of a fact. But, if the algorithm were employed consistently across all the facts of interest and across all articles in a domain, the numbers on the level of a fact would be consistently different and one would still be able to analyze the relative complexity of extracting that fact from a piece of text in the domain.

[1] We wish to thank Jerry Hobbs of SRI for providing us with the rules of their partial parser.

6. Analysis of MUC-4

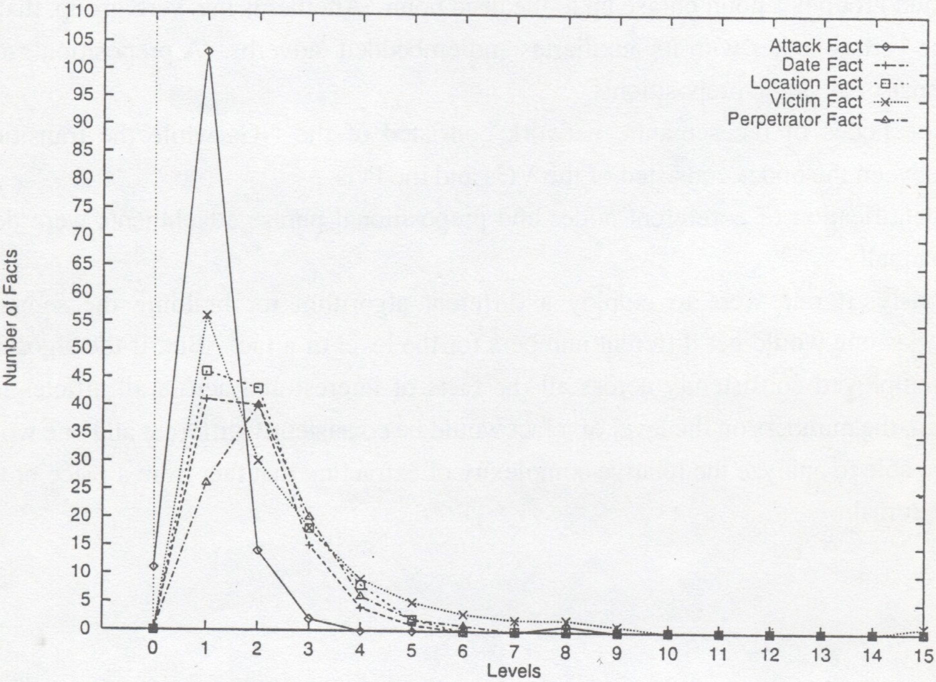

Figure 2 *MUC-4: Level Distribution of Each of the Five Facts*

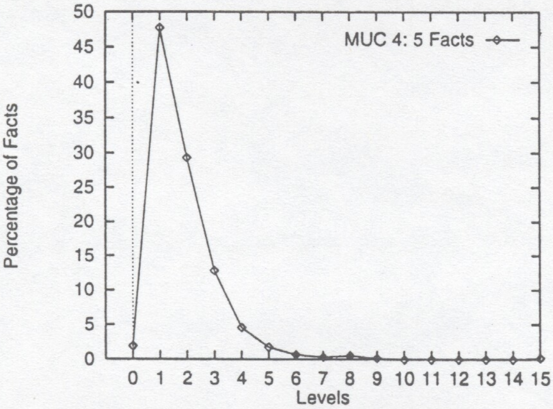

Figure 3 *MUC-4: Level Distribution of the Five Facts Combined*

Based on our definition of the level of a fact, we analyzed the MUC-4 terrorist domain. Based on the official MUC-4 template, we selected a set of *standard facts* that we felt captured most of the information in the template. They are: (The full definition of each fact is not included here.)

•The type of attack.

•The date of the attack.

•The location of the attack.

•The victim (including damage to property).

•The perpetrator(s) (including suspects).

We then built the semantic networks (using the algorithm described in the previous section) for the relevant articles from the MUC-4 TST3 set of 100 articles. From the semantic network for each article, we calculated the levels of each of the five standard facts. The level distribution of the five facts for the MUC-4 TST3 set is shown in Figure 2. The level distribution of the five facts combined is shown in Figure 3.

Based on the data collected above, we made the following observations:

•There were 69 relevant articles in the MUC-4 TST3 set of 100 articles, each reporting one or more terrorist attacks.

•The five facts of interest appeared 570 times in the 69 articles.

•A number of articles reported the same fact at two different places and at two different levels in the same article. The first, usually, in the first paragraph of the text which reported the attack without giving too many details, and, the second, later in the article when the attack was reported with all the details.

As one would expect, the level of the first occurrence of a fact in an article is usually less than or equal to the level of the second occurrence of that fact in the same article.

•From Figure 3, we can see that almost 50% of the five facts were at level-1. This is not surprising because four out of the five *standard* facts most frequently occur as level-1 facts (Figure 2).

6.1 Evaluating the Difficulty of the MUC-4 Terrorist Domain

We extended our analysis to analyze the difficulty of understanding a text in the MUC-4 terrorist domain.

Obviously, the difficulty of understanding a text in a domain depends directly on the expected level of a fact in that domain. We define this expected level of a fact in a domain to be the *domain number* of the domain. The domain number is measured in level units (LUs). Two domains can therefore be compared on the basis of their domain numbers.

The formula used to calculate the domain number is:

$$\frac{\Sigma_{l=0}^{\infty} l * x_l}{\Sigma_{l=0}^{\infty} x_l}$$

where x_l is the number of times one of the *standard facts* appeared at level-*l* in the articles of the domain.

Based on the levels of the five standard facts in the MUC-4 TST3 set of articles, we calculated the domain number of the terrorist domain to be 1.87 LUs. We are assuming the fact that the set of 100 randomly chosen articles in the MUC-4 TST3 set are representative of the domain. This assumption may not necessarily hold, but, given the large number of articles we analyzed, we hope that the domain number calculated is close to the actual domain number of the terrorist domain.

7. Analysis of MUC-5

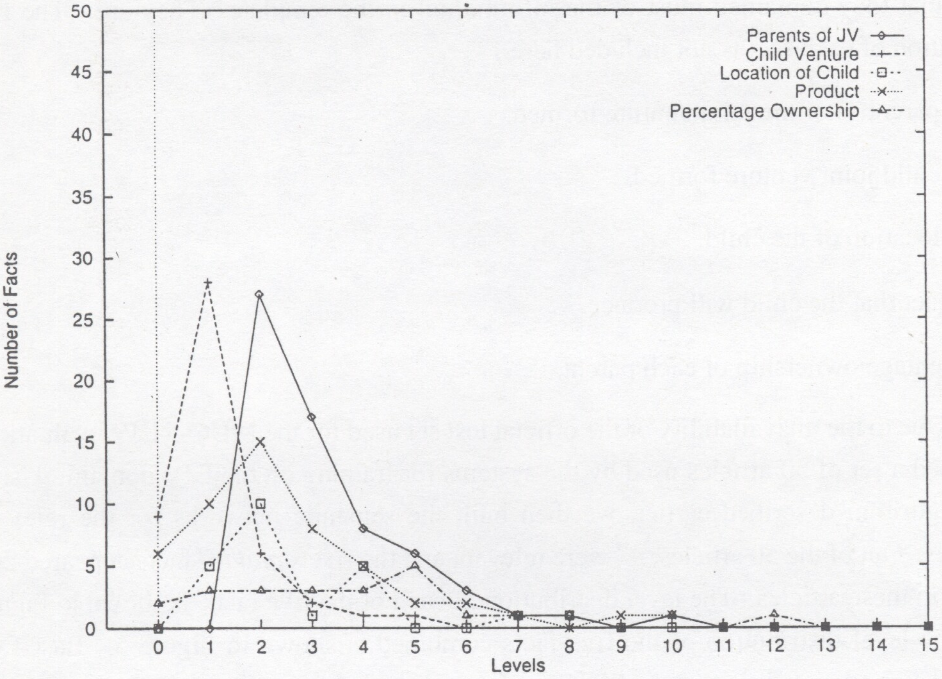

Figure 4 MUC-5: Level Distribution of Each of the Five Facts

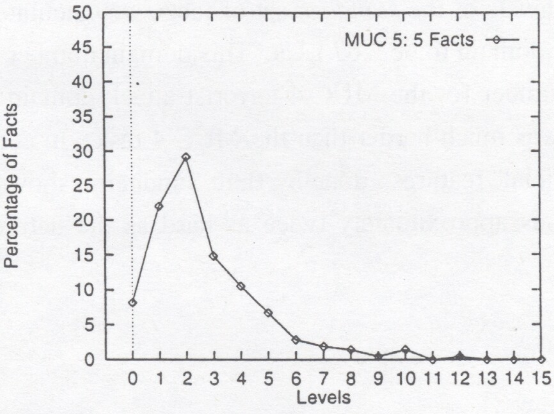

Figure 5 MUC-5: Level Distribution of the Five Facts Combined

Because two different domains were used in MUC-5 (each in two different languages), we decided to focus only on the English Joint Ventures (EJV) domain. Once again, the set of *standard* facts were selected from the official MUC-5 template and were chosen such that they contained most of the information in the template. They are: (The full definition of each fact is not included here.)

- The parent(s) of the joint venture formed.

- The child joint venture formed.

- The location of the child.

- Product that the child will produce.

- Percentage ownership of each parent.

Due to the unavailability of the official test set used for the MUC-5 EJV evaluation, we used a set of 50 articles used by the systems for training on the EJV domain. Using the algorithm described earlier, we then built the semantic networks for the relevant articles. Out of the 50 articles, 47 were relevant and the five *standard* facts appeared 209 times in these articles. The level distribution of each of the five facts is shown in Figure 4. The level distribution of the five facts combined is shown in Figure 5. Based on Figure 4 one can deduce that the MUC-5 EJV domain is harder than the MUC-4 terrorist domain because three out of the five standard facts most frequently occur as level-2 facts. Figure 5 peaks at level-2 giving further indication that the domain number for this domain is more than 2 LUs.

Based on the levels of the *standard* set of facts, we calculated the domain number of the MUC-5 EJV domain to be 2.67 LUs. This domain number is almost 1 LU higher than the domain number for the MUC-4 terrorist attack domain and it shows that the MUC-5 EJV task was much harder than the MUC-4 task. In comparison, an analysis, using more "superficial" features, done by Beth Sundheim, shows that the nature of the MUC-5 EJV task is approximately twice as hard as the nature of the MUC-4 task [Sundheim 93].

8. Analysis of MUC-6

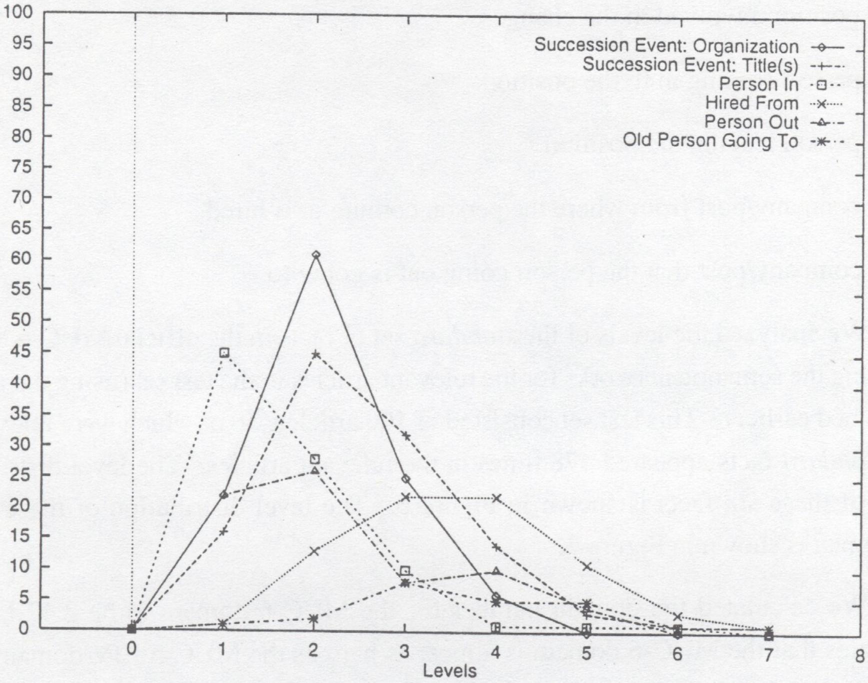

Figure 6 MUC-6: Level Distribution of Each of the Six Facts

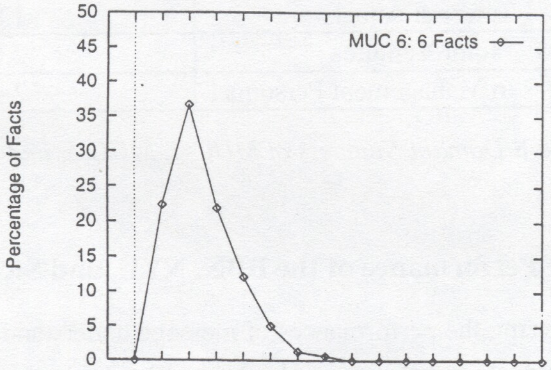

Figure 7 MUC-6: Level Distribution of the Six Facts Combined

The domain used for MUC-6 consisted of articles regarding changes in corporate executive management personnel. As in the case of our analyses of the previous two MUCs, we selected a set of *standard* facts based on the official MUC-6 template. This set consisted of the following facts: (The full definition of each fact is not included here.)

- Organization where the change(s) in the personnel took place.
- The position involved in the changes.
- The person coming in to the position.
- The person leaving the position.
- The company/post from where the person coming in is hired.
- The company/post that the person going out is going to.

We analyzed the levels of the *standard* set of facts in the official MUC-6 test set by building the semantic networks for the relevant articles in the test set (using the algorithm described earlier). This test set consisted of 100 articles, 56 of which were relevant. The six *standard* facts appeared 478 times in the relevant articles. The level distribution of each of these six facts is shown in Figure 6. The level distribution of these six facts combined is shown in Figure 7.

We calculated the domain number for the MUC-6 domain to be 2.47 LUs. This indicates that the MUC-6 domain is almost as hard as the MUC-5 EJV domain. Figure 8 shows the domain numbers for the three MUCs that have been analyzed.

MUC	Domain	Domain Numbers (in LUs)
MUC-4	Terrorist Attacks	1.87
MUC-5	Joint Ventures	2.67
MUC-6	Changes in Management Personnel	2.47

Figure 8 *Domain Numbers of MUC-4, MUC-5, and MUC-6*

9. Analysis of the Performance of the BBN, NYU, and SRI Systems

We felt that by analyzing the performances of message understanding systems based on their ability to extract facts at different levels, we could provide a *deeper* analysis of these systems. Therefore, we decided to look at the templates produced by the BBN, NYU, and SRI systems for the MUC-4 and the MUC-6 tasks. For each of the MUC-4 and MUC-6 domains, we studied these output templates and then examined the performance of each system as it extracted the set of *standard* facts for that domain.

A low performance on level-1 facts certainly points to problems in parsing and basic pattern training for a message understanding system. The main reason being that usually no coreferences have to be resolved when retrieving a level-1 fact. Therefore, when retrieving such a fact, a system only has to recognize patterns in the text. And inability to recognize these patterns points to problems in parsing (assuming that the system has been adapted to the domain well).

On the other hand, a low performance on higher (≥2) level facts points to problems in basic pattern training and the coreferencing module. As mentioned earlier, a level-*k* fact is a union of *k* level-1 facts. Therefore, when retrieving such a fact, a system has to identify each of the *k* components and then the coreferencing module has to piece these *k* facts together.

Although many of the system developers do analyze the performance of their system after a formal evaluation (such as a MUC), our approach provides a more structured basis for doing such an analysis. The added advantage is that the approach is system independent and can be used to gain some insights into systems that developers are not familiar with.

9.1 Limitations of the Analysis

Since we only had access to the final templates produced by the three systems, our analysis was limited by the information present in the templates.

Most systems, when processing an article, produce a set of intermediate templates which are then merged to form one or more final templates for the article. Systems which employ a greedy merging algorithm for merging the intermediate templates generally extract less information from the articles (because they produce a fewer number of templates). On the other hand, systems which employ a lazy merging algorithm for merging the intermediate templates get more information (at the expense of precision) all of which may not be desired.

Systems employing a greedy merging algorithm, naturally, have lower recall compared to systems employing a lazy merging algorithm. And since we had access only to the final templates produced by the systems, these systems had a lower recall in our analysis as well.

Our analysis was also limited by our knowledge of the inner workings of the systems we were analyzing. Since we were not the developers of these systems, we only had access to limited information on the workings of the systems. Most of our information about the systems was derived from the descriptions of the systems found in the proceedings of the Message Understanding Conferences. Therefore, the conclusions that we could draw from the performances of these systems, although insightful, were of a high level. We were able to corroborate most of our conclusions with the information found in the MUC proceedings.

9.2 Analysis Based on MUC-4

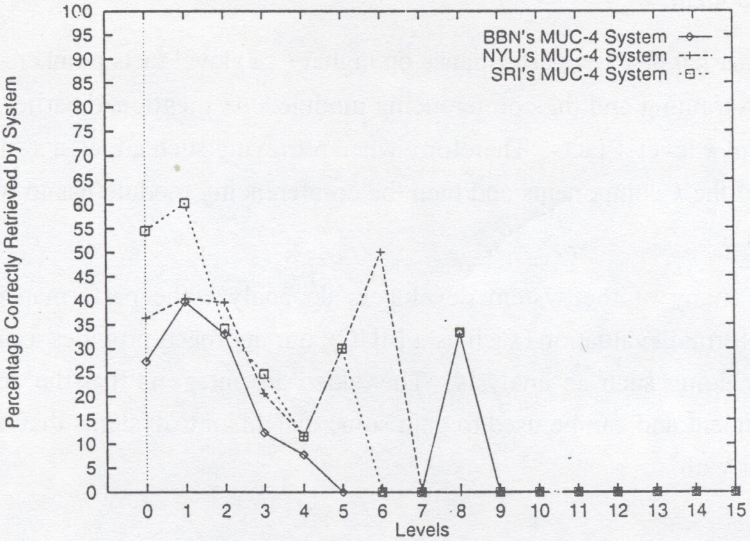

Figure 9 *Performance of the three systems at MUC-4*

System	Actual Recall Achieved in MUC-4	Recall Based on 5 Facts
BBN	30	31
NYU	41	36
SRI	44	44

Figure 10 *MUC-4: Performance of the Three System*

We analyzed the templates produced by the BBN, NYU, and SRI systems for the MUC-4 TST3 set of articles. We then examined the performance of each system based on the five facts of interest. The performance of the three systems across the different levels of the five facts is shown in Figure 9. The significance of the data diminishes greatly for levels bigger than 5 because of the sparsity in the occurrence of these facts. Figure 10 shows the actual recall performance of the three systems on the MUC-4 TST3 set [MUC-4 1992], and the recall performance of the systems based on the five facts (and their levels).

The following observations can be made about the performance of the three systems:

- The recall of the system based on the five facts and their levels is very close to the actual recall achieved by the systems (Figure 10).
- One possible explanation for the relatively large difference in the two recall statistics for

the NYU system is the large number of partially correct answers (PAR) produced by it. When we graded the systems, we did not give any partial credit to a system for getting only a part of a fact (particularly in the case of conjunctions).

9.2.1 BBN's System

BBN's system achieved a recall of 31% on the MUC-4 TST3 task. In addition, the system retrieved about 40% of level-1 facts (Figure 9). Moreover, for higher (≥2) level facts, the system did worse than both the other systems.

A low overall recall did indicate that the system employed a greedy merging algorithm. This indeed was the case [Weischedel 1992]. But, the fact that the system was only able to retrieve only 40% of all level-1 facts did indicate problems with parsing.[2] This was puzzling initially because BBN's system used a Fast Partial Parser (FPP) [Ayuso 1992]; and there are generally fewer problems with partial parsing than with full parsing. A closer analysis of [Weischedel 1992], however, revealed two weaknesses of the system:

- The system had problems with the grammar.
- It was a challenge for discourse processing to be able to collect human targets across sentences.

The first weakness verifies the conclusions drawn from the performance of the system on level-1 facts because problems with the grammar do spill over into parsing. Since, the human target fact was the second most commonly occuring fact (Figure 2), the second weakness (coupled with the first weakness) verifies that the system had some problems with discourse processing; thereby verifying the conclusions drawn from the performance of the system on higher level facts.

9.2.2 NYU's System

NYU's system, like BBN's system, achieved a performance of around 40% on level-1 facts. However, unlike BBN's system, NYU's system performed relatively better on higher level facts.

NYU's MUC-4 system attempted to generate a full parse of the sentences. Since the MUC-4 corpus contained articles that had been translated from news broadcasts, the

[2] We assume here that all the MUC-4 systems were trained well. This is because all theparticipants were given a training corpus consisting of 1300 messages well in advance of the final evaluations. This assumption does not hold for MUC-6 because the participants were given a training set of 100 messages only a month before the final evaluations.

articles contained missing (indistinct) words and words with spelling errors. All of these factors led to problems in parsing for the system [Grishman 1992]. This agrees with the conclusions drawn from the performance of the system on level-1 facts.

Despite a relatively poor performance on level-1 facts, NYU's system performed (relatively) better on higher level facts. Since the performance on level-1 facts does affect the performance on higher level facts, we conclude that NYU's coreferencing module performed better than the coreferencing modules of the other two systems. This fact is actually verified later in this paper.

9.2.3 SRI's System

SRI's system had an overall recall of 44%. The system performed extremely well on level-1 facts extracting about 60% of these facts. In addition, the system's performance on higher level facts was comparable to the performance of NYU's system on such facts.

SRI's system used a conservative (lazy) merging strategy for template merging. In addition, the system used a partial parser that achieved a precision of 96.4% [Hobbs 1992]; which accounts for the good performance on level-1 facts. Given the high performance on level-1 facts, we expected the system to perform better on higher level facts. A closer analysis of [Hobbs 1992] revealed, however, that the system used a very simple coreferencing strategy which included a "rudimentary sort of pronoun resolution." This coupled with the lazy merging strategy accounts for the performance on higher level facts.

9.3 Analysis Based on MUC-6

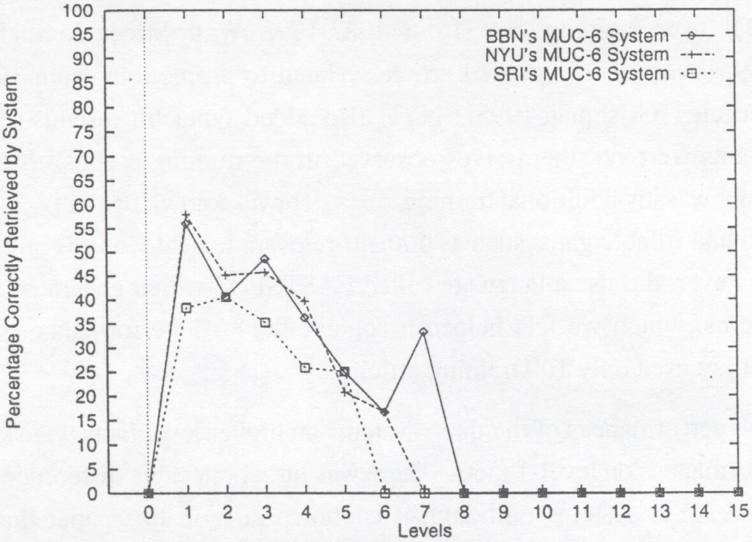

Figure 11 *Performance of the three systems at MUC-6*

We continued our analysis by examining the templates produced by the BBN, NYU, and SRI systems for the MUC-6 test set. The performance of the three systems on the six facts is shown in Figure 11. As in the case of MUC-4, the significance of the data is significantly reduced for levels bigger than 5.

9.3.1 Analysis of the Three Systems

The corpus used for MUC-6 consisted of articles from the Wall Street Journal. This eliminated some of the problems including missing (indistinct) words, mis-spelled words, and grammatical errors that the systems had to deal with in the MUC-4 corpus. Also NYU's system was moved from using full parsing to partial parsing. This meant that all the three systems used partial parsing for MUC-6. None of the three systems reported any problems with parsing.

But, the performances of the systems differed greatly. For MUC-6, the BBN and NYU systems retrieved around 57% of level-1 facts while SRI extracted around 40%. This was in stark contrast to MUC-4 where the roles were reversed with SRI's system extracting around 60% of level-1 facts and BBN's and NYU's systems extracting around 40%. Since none of the systems reported any problems with parsing, the relatively large difference in the performances of the three systems on level-1 facts certainly pointed to differences in the quality of training.

The official MUC-6 training set consisted of only 100 articles which were given to the participating groups a month before the final evaluation. Because the amount of training data was so little, both BBN and NYU decided to get additional training materials. BBN used training data from 1987-1992 Wall Street Journal articles [Weischedel 1995] while NYU studied articles related to promotions from 1987 Wall Street Journal articles [Grishman 1995]. NYU also added syntactic variants of patterns even if the variants were not themselves observed in the training corpus. SRI, on the other hand, did not use any additional training texts. They acknowledge in [Appelt 1995] that their system had fillable gaps such as domain-relevant lexical features on important words. SRI, however, did use a language called FASTSPEC which generated syntactic variants of patterns; which we feel helped it achieve the 40% performance on level-1 facts given that they used only 100 training articles.

The relative performances of the three systems on higher level facts were similar to the relative performances on level-1 facts. There was little noticeable difference between the BBN and the NYU systems (although it is shown later in this paper that NYU's system performed better than BBN's system on higher level facts).. SRI's system, because of its performance on level-1 facts, performed worse than the other two systems on higher level facts.

10. The Role of Coreferencing

We extended our work by further analyzing the performances of the three systems. The new analysis done was motivated by the Beth Sundheim's exploratory study which was mentioned earlier [Hirschman 1992].

Since we had already done an analysis regarding the levels of facts (the distribution of information in a message) and their effect on the performance of message understanding systems, we decided to also look at the the effect of discourse processing, specifically coreferencing, on the performance of message understanding systems. We decided, for each level, to calculate the number of coreferent nodes that comprised facts at that level. We also wanted to analyze the performances of message understanding systems based on the number of coreferences present in the facts retrieved by such a system. As before, the analysis was using data from MUC-4 and MUC-6.

10.1 Analysis of MUC-4

For each *standard* fact at a particular level, we calculated the number of coreferent nodes that comprised the fact at that level. Figure 12 shows, for each level, the number of coreferences for all the *standard* facts at that level. Figure 13 shows the number of coreferences for all the levels combined. Because of data sparsity, the significance of the data diminishes greatly for the number of coreferences ≥ 2.

A closer look at the curves for each level in Figure 12 shows that as the level number increases, the percentage of facts having a larger number of coreferent nodes increases. For example, the curves for levels 0, 1, 2, and 3 peak when the number of coreferences equal 0, the curves for levels 4, 5, and 6 peak when the number of coreferences equal 1, and the curve for level 7 peaks when the number of coreferences equal 2. This is to be intuitively expected.

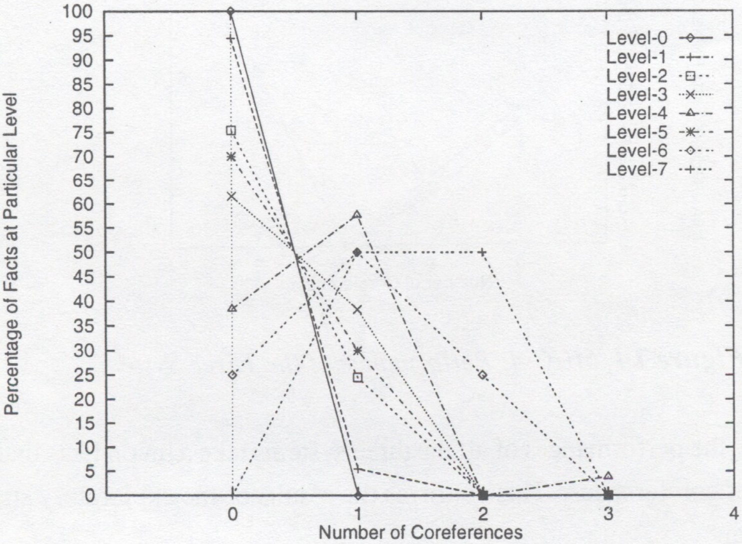

Figure 12 *MUC-4: Number of Coreferences At Each Level*

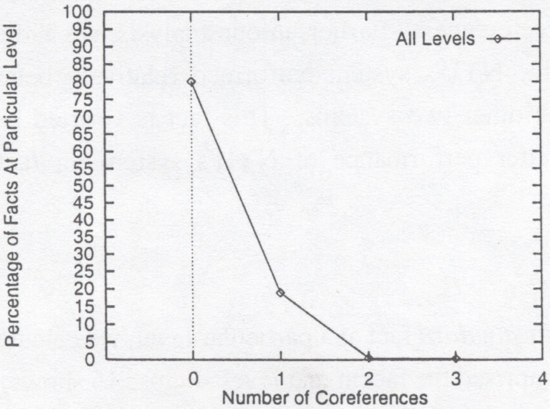

Figure 13 *MUC-4: Number of Coreferences At All Levels*

10.2 Analysis of the Three Systems

We analyzed the performances of the three systems on the standard facts. The performances of the three systems for all levels is shown in Figure 14.

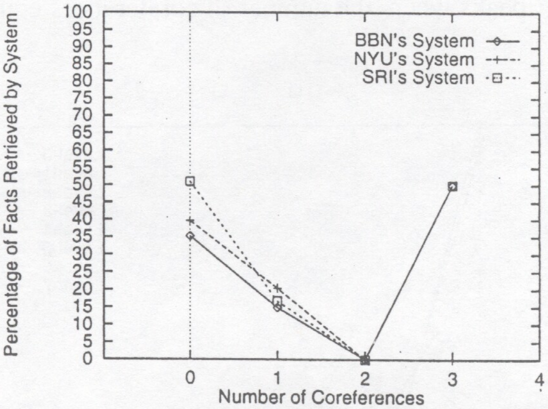

Figure 14 *MUC-4: Performance of the Three System*

As expected, the performances of all the three systems take a hit on facts that contain a larger number of coreferences. This confirms the results of the exploratory study done by Beth Sundheim.

The performances of the three systems on facts that had no coreferences is almost the same as their performances on level-1 facts. This is not surprising at all since most level-1 facts have no coreferences. Earlier, in our analysis, we had concluded that the coreferencing module for NYU's system performed relatively better than the coreferencing modules of the other two systems. This fact is verified in Figure 14 which shows the relatively better performance of NYU's system on facts containing one coreferent node.

10.3 Analysis of MUC-6

As with MUC-4, for each *standard* fact at a particular level, we calculated the number of coreferent nodes that comprised the fact at that level. Figure 15 shows, for each level, the number of coreferences for all the *standard* facts at that level. Figure 16 shows the number of coreferences for all the leveis combined. Because of data sparsity, the significance of the data diminishes greatly for the the number of coreferences ≥ 3.

Once again, a closer look at the curves for each level in Figure 15 shows that as the level number increases, the percentage of facts having a larger number of coreferent nodes increases (the curves for levels 1 and 2 peak when the number of coreferences equal 0, the curves for levels 3, 4, and 5 peak when the number of coreferences equal 1, and the curve for level 6 peaks when the number of coreferences equal 2).

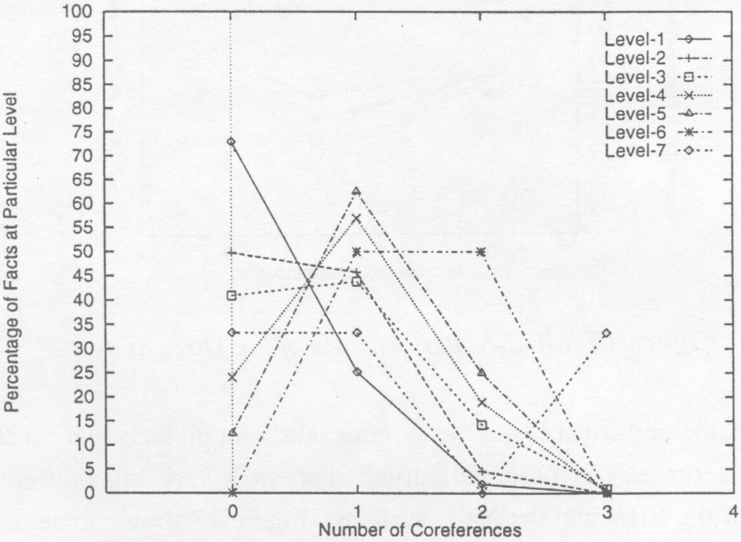

Figure 15 MUC-6: Number of Coreferences At Each Level

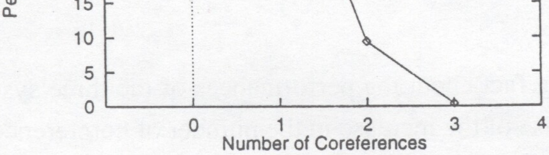

Figure 16 MUC-6: Number of Coreferences At All Levels

10.4 Analysis of The Three Systems

We analyzed the performance of the three systems on the standard facts. The performances of the three systems for all levels is shown in Figure 17.

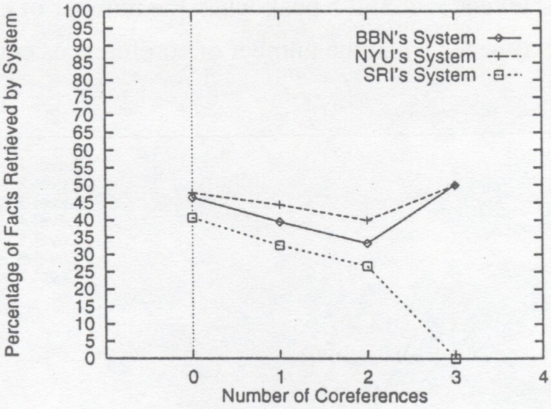

Figure 17 *MUC-6: Performance of the Three System*

As before, the performances of the systems take a hit on facts that contain a larger number of coreferences. As observed earlier, there was very little difference in the performances of the BBN and the NYU systems. Figure 17 shows some of the differences. In particular, it shows the relatively better performance of NYU's system on facts containing larger number of coreferences.

Comparing Figure 14 with Figure 17 one can see that the performances of the systems on facts containing larger number of coreferences has improved considerably since MUC-4. This is a result of realization of the importance of discourse processing. It is also the result of a conscious effort on the part of the people organizing the MUCs to get the groups developing the systems to focus on discourse processing (specifically coreferencing). Coreferencing was introduced as a formal (although optional) task in MUC-6. And a number of groups undertook efforts to specifically improve their coreferencing modules.

But, the surprising fact about the performances of the three systems for MUC-6 is that the hit taken because of the increase in the number of coreferences is approximately the same (Figure 17). This shows that while improvements in the coreferencing modules have helped the systems perform better, the improvements have been almost the same for the three systems. The basic difference in the performances of the three systems has stemmed mainly from their performances on level-1 facts (facts with almost no coreferences). Therefore, for information extraction systems to achieve recall and precision of 70% or higher, there has to be significant improvements in their ability to process discourse.

11. Future Work

We are currently looking at the possibility of converting this method of analyzing the performance of message understanding systems to a method for predicting the performance of such systems on a particular domain. One obvious way of being able to predict the performance of a system on a particular domain is as follows: First calculate the level distribution of a set of standard facts for the domain. And then, based on the past performances of the system, at each level, calculate the expected performance.

12. Conclusions

In this paper we introduce a new method of classifying a fact based on the degree of difficulty of extracting it from text. This classification mechanism is then used to analyze the degree of difficulty of understanding a text in a domain. This analysis is a big step up from some of the methods used earlier. The classification mechanism is also used to analyze the performance of three message understanding systems on two different MUC domains. The analysis is then extended to examine the role of coreferencing in the performance of these systems.

In addition to providing a *deeper* insight into the performances of the three systems, our analysis has also brought out the following two points:

(1) As seen in the performances of the systems on MUC-6, the amount of training done on a message understanding system is important. And, therefore, being able to predict the amount of training needed to port a system to a particular domain is an area that needs attention.

(2) While considerable improvements have been made in the amount and the quality of discourse processing (particularly coreferencing) done by message understanding systems, a lot more needs to be done for the systems to be able to break the 65-70% overall performance barrier.

Acknowledgments

The authors thank Beth Sundheim for providing them with the official data from the MUCs.

The first author also acknowledges IBM Corporation for providing him with a Fellowship under which this work was done.

References

Appelt, Douglas E., et al. "SRI International: Description of the FASTUS System Used for MUC-6," *Proceedings of the Sixth Message Understanding Conference* (*MUC-6*), 1995, pp.

237-248.

Ayuso, D., et al. "BBN: Description of the PLUM System as Used for MUC-4," *Proceedings of the Fourth Message Understanding Conference* (*MUC-4*), 1992, pp. 169-176.

Grishman, Ralph. "New York University: Description of the PROTEUS System as Used for MUC-4," *Proceedings of the Fourth Message Understanding Conference* (*MUC-4*), 1992, pp. 233-241.

Grishman, Ralph. "The NYU System for MUC-6 or Where's the Syntax?" *Proceedings of the Sixth Message Understanding Conference* (MUC-6), 1995, pp. 167-175.

Hendrix, Gray G. "Encoding Knowledge in Partitioned Networks." In *Associative Networks*. Nicholas V. Findler (ed.). New York: Academic Press, 1979, pp. 51-92.

Hirschman, Lynette. "An Adjunct Test for Discourse Processing in MUC-4," *Proceedings of the Fourth Message Understanding Conference* (*MUC-4*), 1992, pp. 67-77.

Hobbs, J., et al. "SRI International: Description of the FASTUS System Used for MUC-4," *Proceedings of the Fourth Message Understanding Conference* (*MUC-4*), 1992, pp. 268-275.

Proceedings of the Third Message Understanding Conference (*MUC-3*), 1991, San Mateo: Morgan Kaufmann.

Proceedings of the Fourth Message Understanding Conference (*MUC-4*), 1992, San Mateo: Morgan Kaufmann.

Proceedings of the Sixth Message Understanding Conference (*MUC-6*), 1995, San Francisco: Morgan Kaufmann.

Schubert, Lenhart K., et. al. "The Structure and Organization of a Semantic Net for Comprehension and Inference." In *Associative Networks*. Nicholas V. Findler (ed.). New York: Academic Press, 1979, pp. 121-175.

Sundheim, Beth M. "TIPSTER/MUC-5 Information Extraction System Evaluation," *Proceedings of the Fifth Message Understanding Conference* (*MUC-5*), 1993, pp. 27-44.

Sundheim, Beth M. "Overview of Results of the MUC-6 Evaluation," *Proceedings of the Sixth Message Understanding Conference* (*MUC-6*),1995, pp. 13-31.

Weischedel, R., et al. "BBN PLUM: MUC-4 Test Results and Analysis," *Proceedings of the Fourth Message Understanding Conference* (*MUC-4*), 1992, pp. 87-94.

Weischedel, Ralph. "BBN: Description of the PLUM System as Used for MUC-6," *Proceedings of the Sixth Message Understanding Conference* (*MUC-6*),1995, pp. 55-69.