Analyzing the Complexity of a Domain With Respect To An Information Extraction Task

Amit Bagga*
Alan W. Biermann
Dept. of Computer Science
Box 90129, Duke University
Durham, N. C. 27708–0129. USA
Internet: {amit, awb}@cs.duke.edu

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 also 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. In addition, we undertake two studies. The first study evaluates the effect of levels on the performance of message understanding systems while the second evaluates the effect of discourse processing, specifically coreferencing, on the performance of message understanding systems.

*Supported by a fellowship from IBM Corporation.

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 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. Moreover, we also 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. Based on our classification mechanism, we undertake two studies. The first one evaluates the the performance of three MUC systems (BBN, NYU, and SRI) based on their ability to extract a set of "standard" facts (at different levels) from the MUC-4 terrorist reports domain. The second study evaluates the effect of discourse processing, specifically coreferencing, on the performance of the three systems.

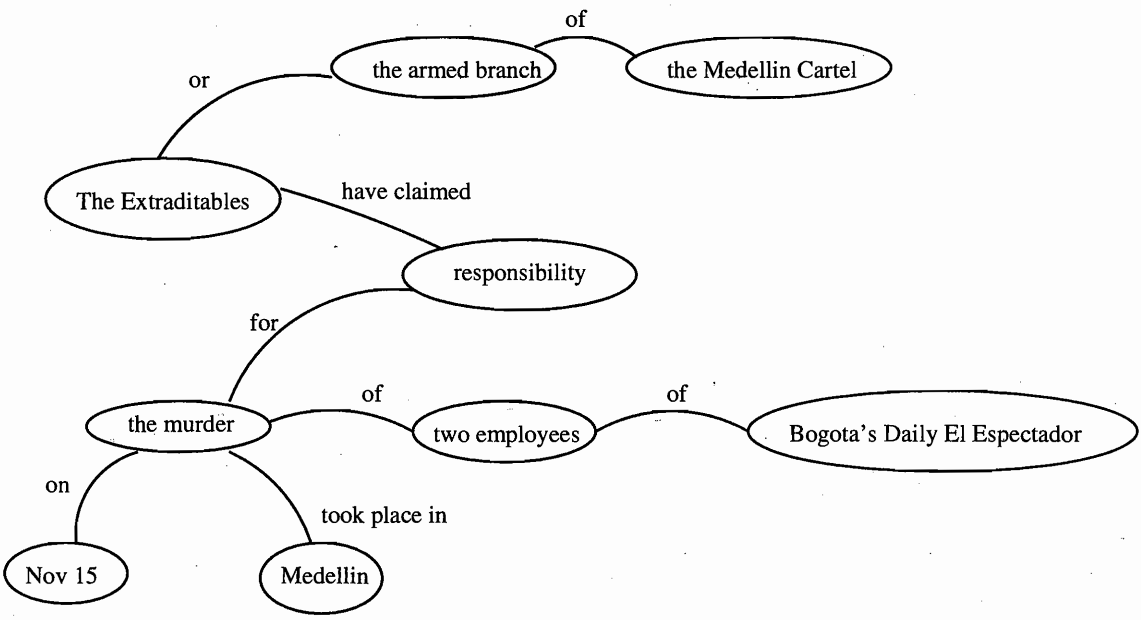

Figure 1: A Sample Semantic Network

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 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.

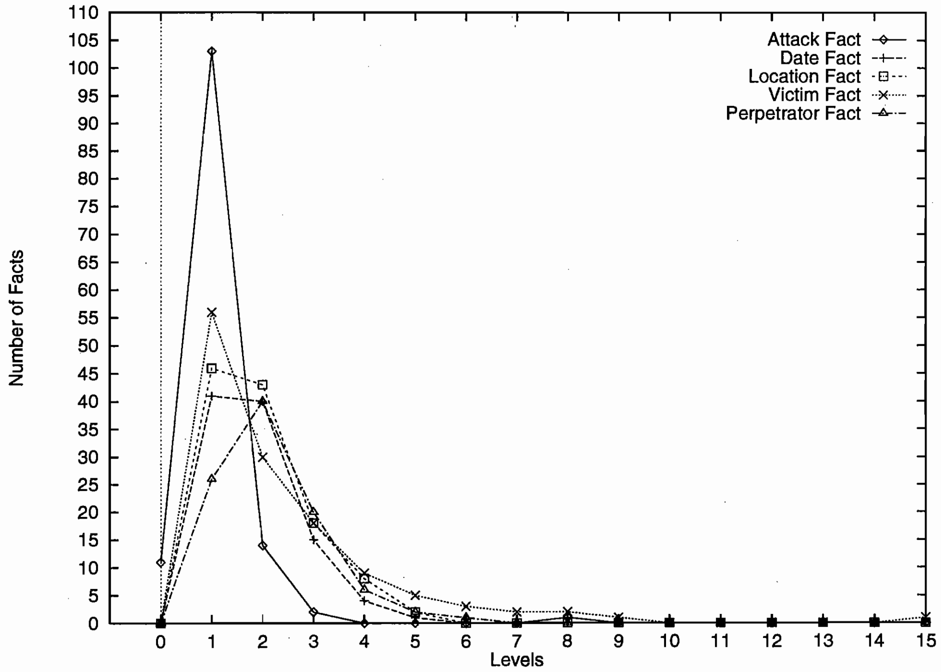

Figure 2: MUC-4: Level Distribution of Each of the Five Facts

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 $\{x_1, x_2, \ldots, x_n\}$. Let s be the partial semantic network consisting of the set of nodes $\{x_1, x_2, \ldots, x_n\}$ interconnected by the set of arcs $\{t_1, t_2, \ldots, t_k\}$.

   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, 1992).

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, 1992), 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. Our analysis shows that all the three systems consistently did much worse on higher level facts. In addition to confirming Hirschman's hypothesis, the analysis also shows that higher level facts are indeed harder to extract. Some details of the analysis are given later in this paper. (Bagga, 1997) gives the complete details about the analysis.

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].

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.

6 Analysis of MUC-4

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.

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

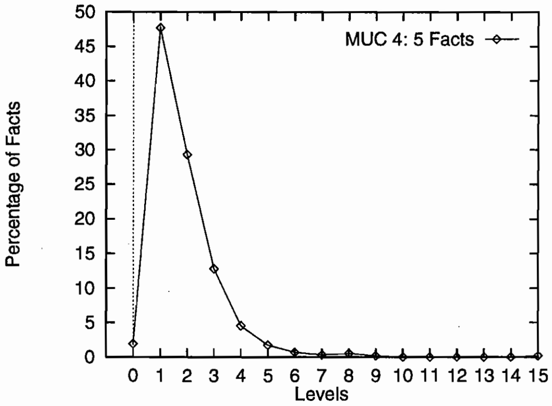

Figure 3: MUC-4: Level Distribution of the Five Facts Combined

- 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:

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

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

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.

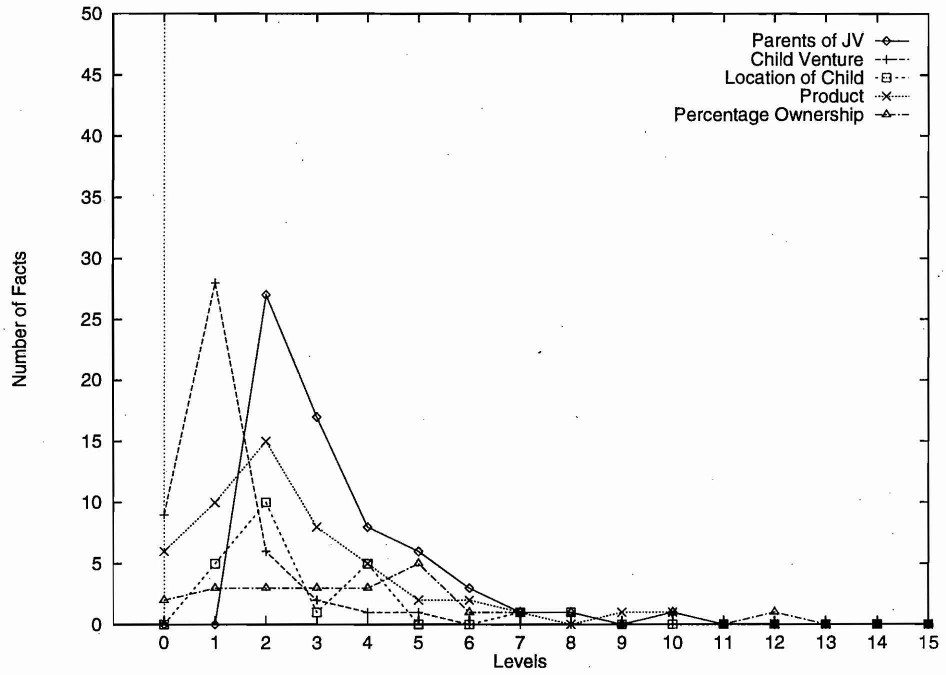

Figure 4: MUC-5: Level Distribution of Each of the Five Facts

- 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

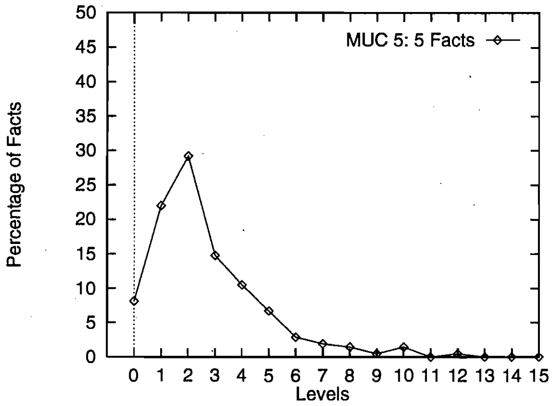

Figure 5: MUC-5: Level Distribution of the Five Facts Combined

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, 1993).

8 Analysis of MUC-6

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

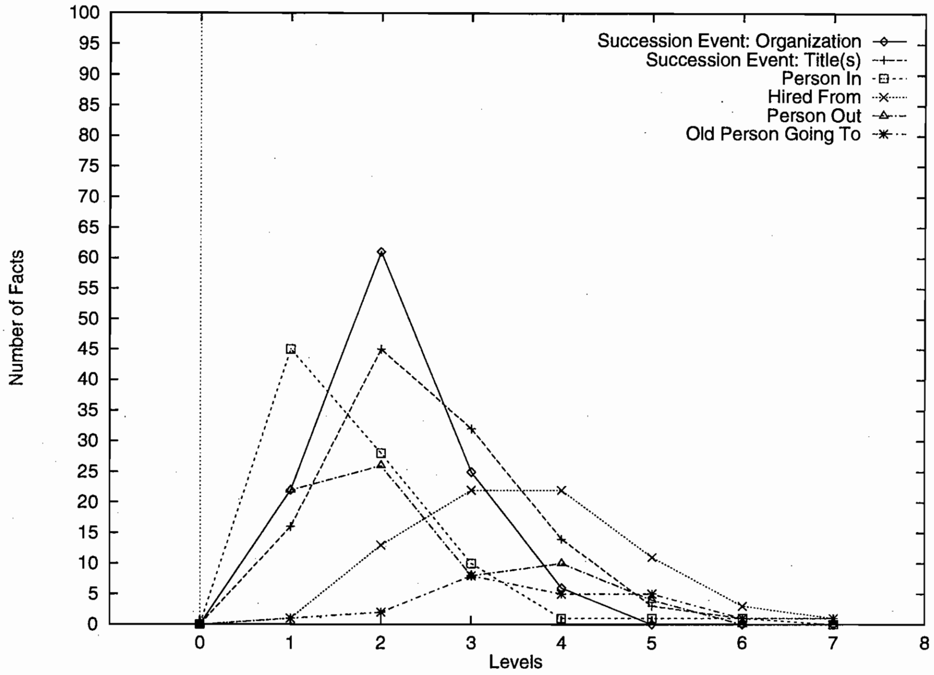

Figure 6: MUC-6: Level Distribution of Each of the Six Facts

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.

9 Extending the Analysis

Motivated by the exploratory study done by Beth Sundheim, we decided to undertake two studies. The first one was to do an analysis regarding the levels of facts (the distribution of information in a message) and their effect on the performance of message understanding systems. The second study was to look at the the effect of discourse processing, specifically coreferencing, on the performance of message understanding systems.

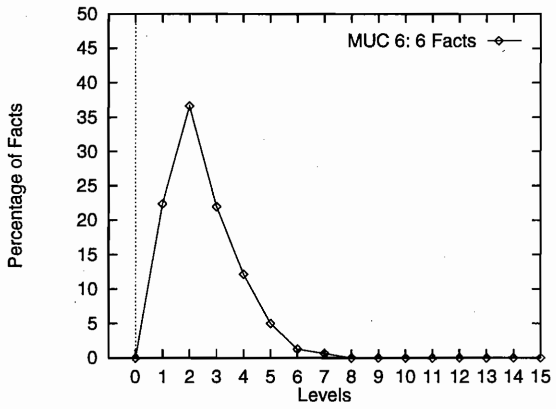

Figure 7: MUC-6: Level Distribution of the Six Facts Combined

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

9.1 Analysis of the Performance of Information Extraction Systems

We continued our analysis by examining the templates produced by the BBN, NYU, and SRI systems for the MUC-4 TST3 set of articles. We studied each template and then examined the performance of each system as it extracted the five *standard* facts for the domain. The performance of the three systems across the different levels of the five facts is shown in Figures 9, 10, and 11. The figures show the degradation in the performance of all the three systems on higher level facts. The significance of the data diminishes greatly for levels bigger than 4 because of the sparsity in the occurrence of these facts.

This type of analysis forms the basis for providing greater insight into the performances of information extraction systems. For example, 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).

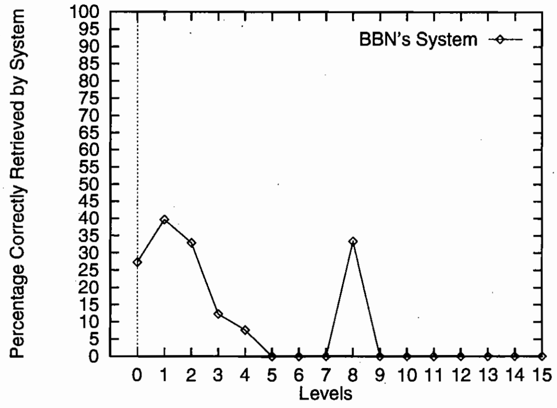

Figure 9: Performance of BBN's MUC-4 System

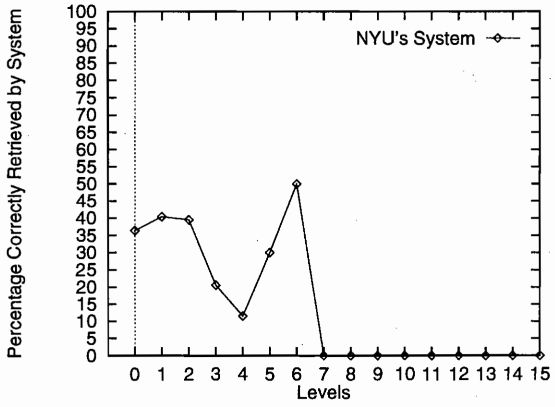

Figure 10: Performance of NYU's MUC-4 System

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.

More details on such an analysis can be found in (Bagga, 1997).

9.2 The Role of Coreferencing

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. The analysis was using data from MUC-4 and MUC-6.

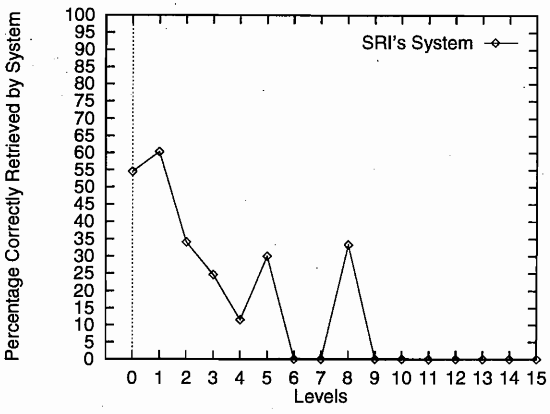

Figure 11: Performance of SRI's MUC-4 System

9.2.1 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.

9.2.2 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.

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. Moreover, 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.

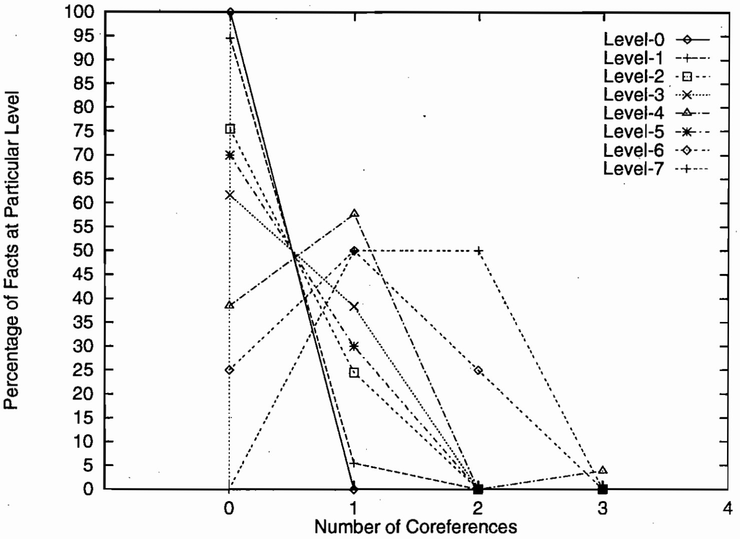

Figure 12: MUC-4: Number of Coreferences At Each Level

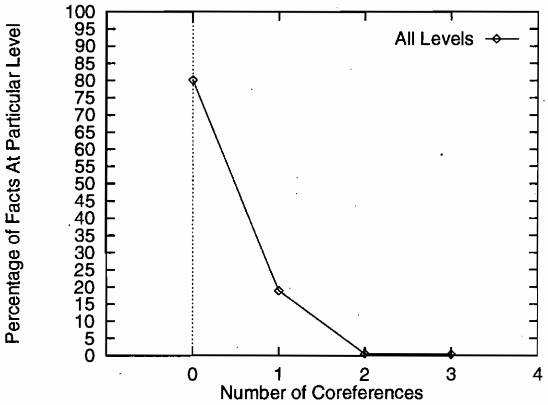

Figure 13: MUC-4: Number of Coreferences At All Levels

9.2.3 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 levels 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

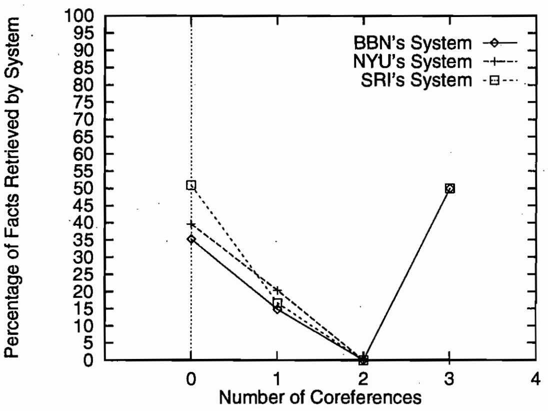

Figure 14: MUC-4: Performance of the Three System

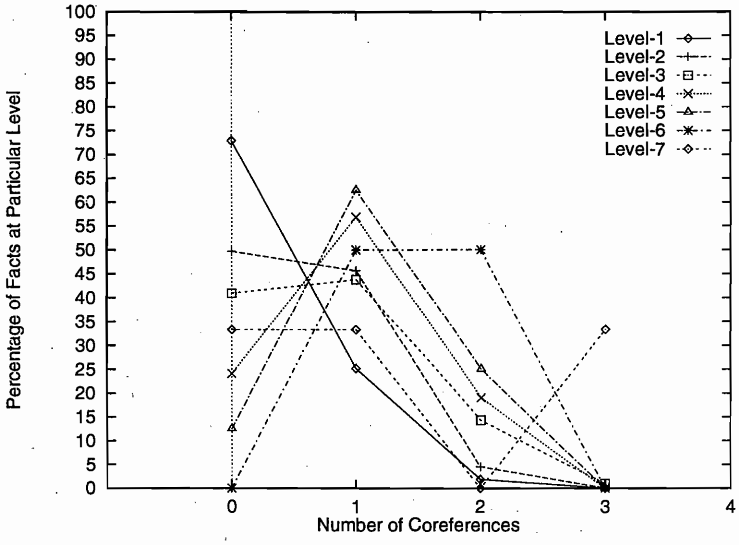

Figure 15: MUC-6: Number of Coreferences At Each Level

for level 6 peaks when the number of coreferences equal 2).

9.2.4 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. As before, the performances of the systems take a hit on facts that contain a 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).

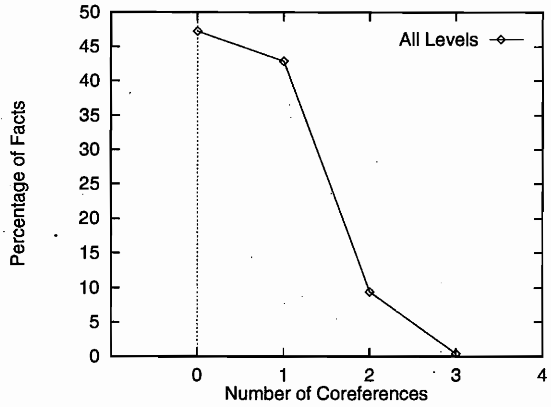

Figure 16: MUC-6: Number of Coreferences At All Levels

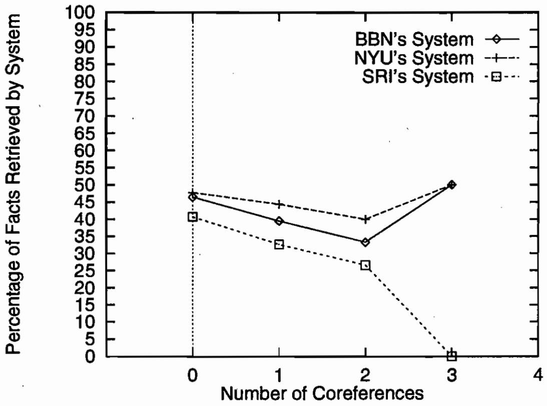

Figure 17: MUC-6: Performance of the Three System

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.

10 Conclusion

The level of a fact with respect to a semantic network for a piece of text provides a new method of classifying a fact based on the degree of difficulty of extracting it from that text. The analysis of the degree of difficulty of understanding a text in a domain comes as a by-product of our approach and is a big step up from some of the techniques used earlier.

11 Acknowledgments

We wish to thank Beth Sundheim for providing us with the official data from the MUCs.

References

Bagga, Amit. Analyzing the Performance of Message Understanding Systems, To Appear.

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)*, pp. 67-77, June 1992.

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

Proceedings of the Sixth Message Understanding Conference (MUC-6), November 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)*, pp. 27-44, August 1993.

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