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

Amit Bagga
Dept. of Computer Science
Box 90129, Duke University
Durham, N. C. 27708–0129
Email: amit@cs.duke.edu
Phone: 919-660-6507

Abstract

In this paper we describe a method of classifying facts (information) into categories or levels; where each level signifies a different degree of syntactic complexity related to a fact. 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.

Introduction

The Message Understanding Conferences (MUCs) have been held with the goal of qualitatively evaluating message understanding systems. While the MUCs held thus far have been quite successful at providing such an evaluation, 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 syntactic complexity related to a fact. Based upon 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, using the proposed classification mechanism, we analyze the complexity of the MUC-7 information extraction task and compare it to the complexities of the information extraction tasks of MUC-4, MUC-5, and MUC-6.

Definitions

Network:
A *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 Network:
A *partial network* is a collection of nodes interconnected by an accompanying set of arcs where the collection

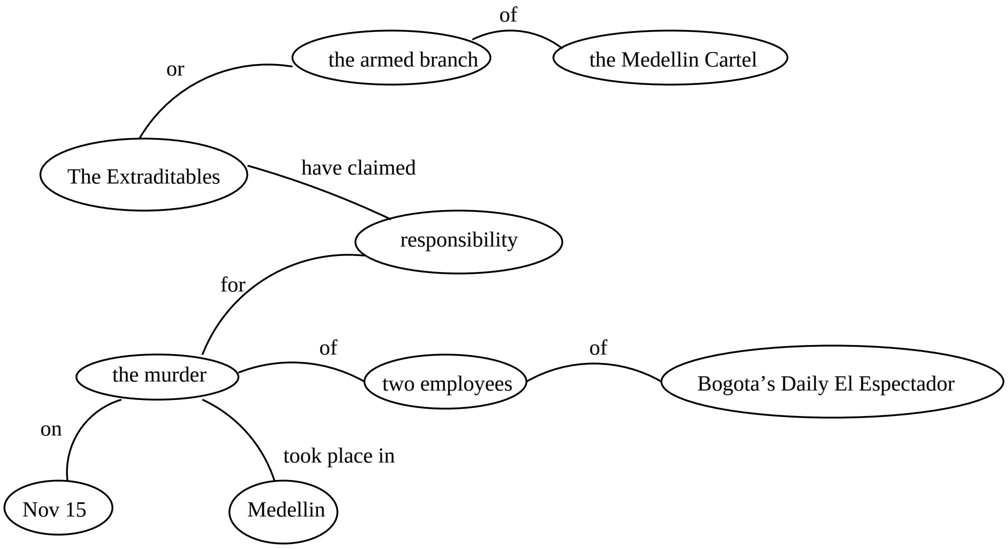

Figure 1: A Sample Network

of nodes is a subset of a collection of nodes forming a network, and the accompanying set of arcs is a subset of the set of arcs accompanying the set of nodes which form the network.

Figure 1 shows a sample 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.

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 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 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 network, S, to be equal to k, the number of arcs linking the nodes which comprise the fact F.

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 network 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 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 network constructed for that piece of text. Therefore, the level of a fact with respect to a 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 network built at the phrase level (i.e. noun groups represent objects while verb groups and preposition groups represent the relationships between the objects).

Examples

Let S be the 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 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 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 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.

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" [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 several information extraction systems on the MUC-4 terrorist reports domain. Our analysis shows that all the 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. [Bagga 1998] gives the complete details about the analysis.

Building the Networks

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

For the set of experiments described in the rest of the paper we used the following algorithm to build the 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 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 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.

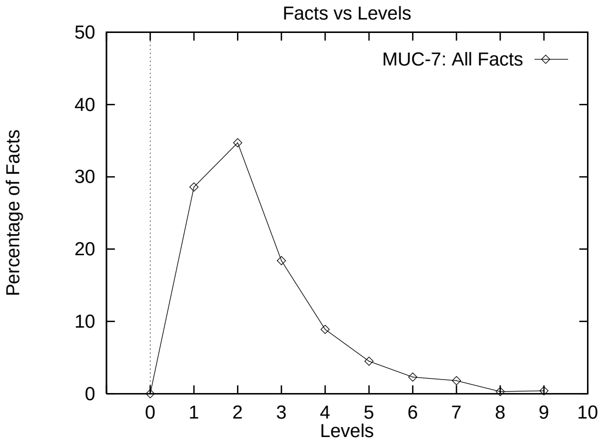

Figure 2: MUC-7: Level Distribution of the Facts Combined

Analysis of the MUC-7 Domain

Based on our definition of the level of a fact, we analyzed the MUC-7 domain which consisted of reports on air vehicle launches. We selected a set of *standard* facts from the official MUC-7 template that we felt captured most of the information in the template. This set consists of:

- Launch Date, Launch Site
- Mission Type, Mission Function, Mission Status
- Vehicle, Vehicle Type, Vehicle Owner, Vehicle Manufacturer
- Payload, Payload Type, Payload Function, Payload Owner
- Payload Manufacturer, Payload Origin, Payload Target, Payload Recipient

Using the algorithm described in the previous section, we built the networks for 32 of the 64 relevant articles in the 100 article MUC-7 test set. From the network for each article, we calculated the level of each of the standard facts mentioned in the article. The standard facts appeared 619 times in the 32 articles analyzed. Figure 2 shows the level distribution of all the standard facts. In addition, Figures 3, 4, 5, and 6 show the level distributions of each of the standard facts.

Figures 4 and 5 show that the curves for the Vehicle and the Vehicle Type facts, and the Payload and Payload Type facts are nearly the same. The main reason being that the phrases (in the text) which describe the vehicles and the payloads, in most cases, also mention their types. For example, the phrase "Long March 3B Rocket" describes the air launch vehicle and also mentions its type.

The Vehicle Owner and Manufacturer facts, and the Payload Owner and Manufacturer facts all occur at higher levels than the Vehicle and the Payload facts indicating that they are harder to extract. The Payload recipient fact appears to be the most complex fact occurring most frequently at level-9 (Figure 6).

Evaluating the Difficulty of the MUC-7 Domain

We extended our analysis to analyze the difficulty of understanding a text in the MUC-7 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.

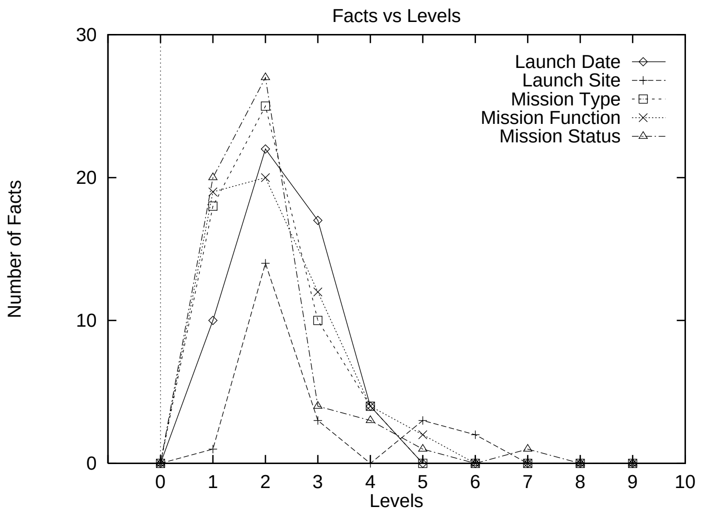

Figure 3: MUC-7: Level Distribution of Each of the Facts

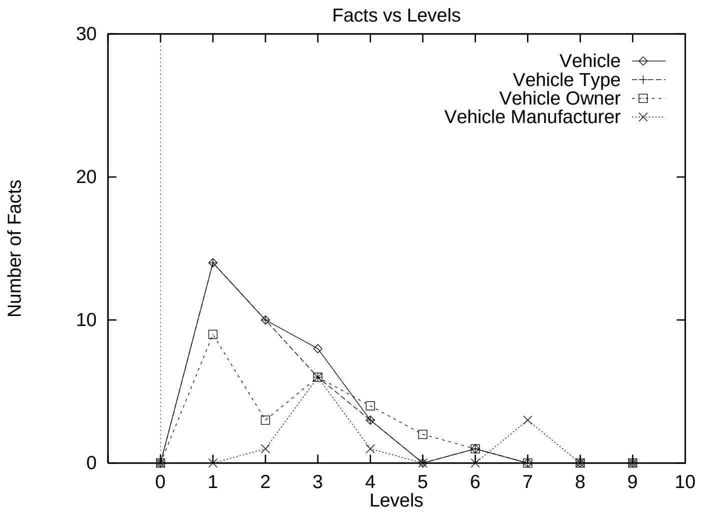

Figure 4: MUC-7: Level Distribution of Each of the Facts

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 standard facts in the MUC-7 test set, we calculated the domain number of the air vehicle launch domain to be 2.44 LUs. The fact that the domain number of this domain is greater than 2 is to be expected given the fact that most curves in Figures 3, 4, 5, and 6 peak at level-2 or higher.

Analysis of MUC-4

The MUC-4 domain consisted of articles reporting terrorist activities in Latin America. 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.)

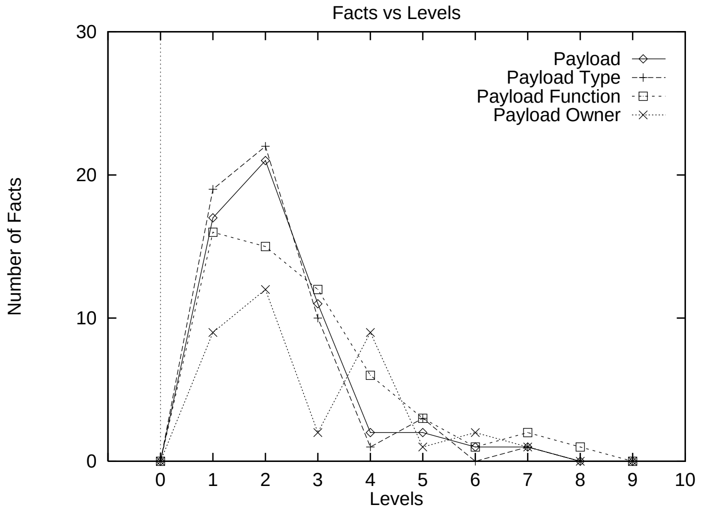

Figure 5: MUC-7: Level Distribution of Each of the Facts

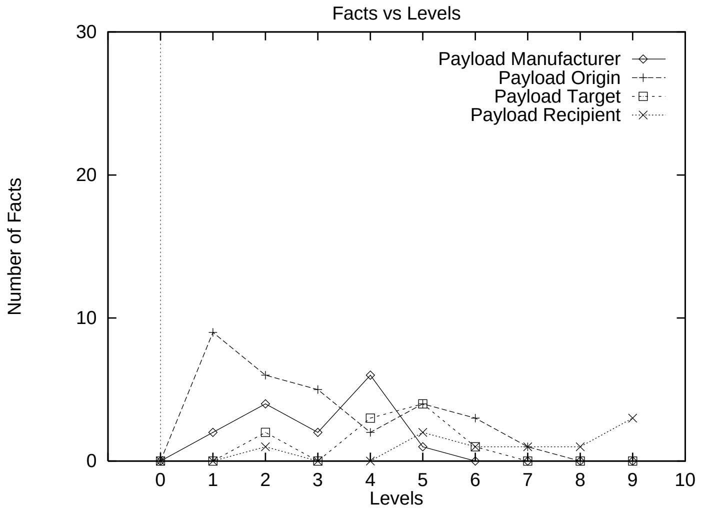

Figure 6: MUC-7: Level Distribution of Each of the Facts

- 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 networks (using the algorithm described earlier) for the relevant articles from the MUC-4 TST3 set of 100 articles. From the 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 7. The level distribution of the five facts combined is shown in Figure 8.

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

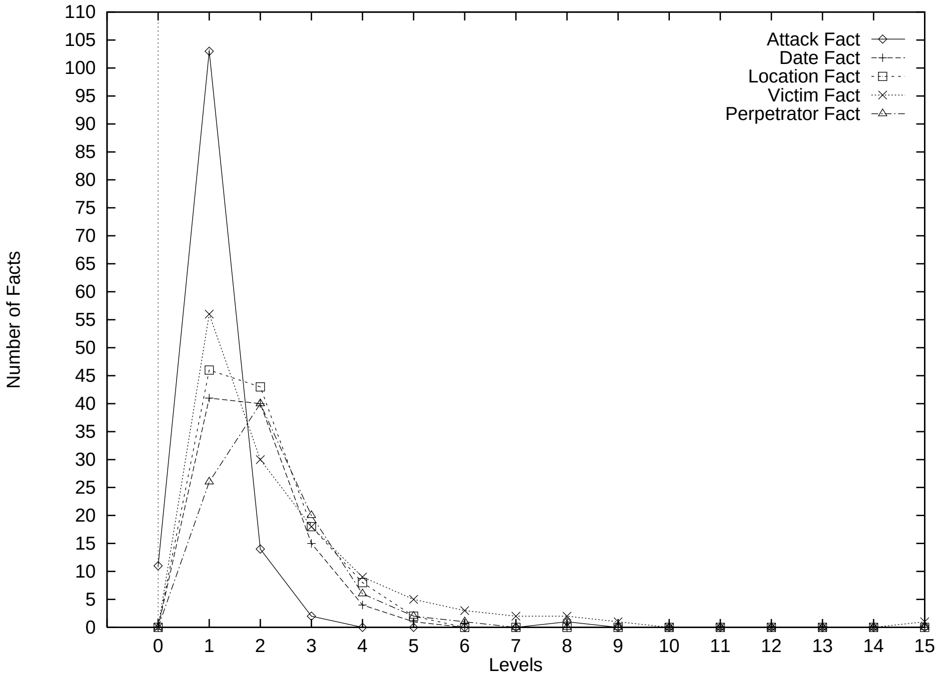

Figure 7: MUC-4: Level Distribution of Each of the Five Facts

- 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 8, 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 7).

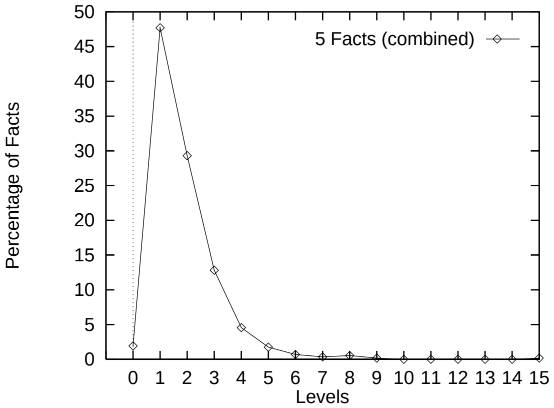

Figure 8: MUC-4: Level Distribution of the Five Facts Combined

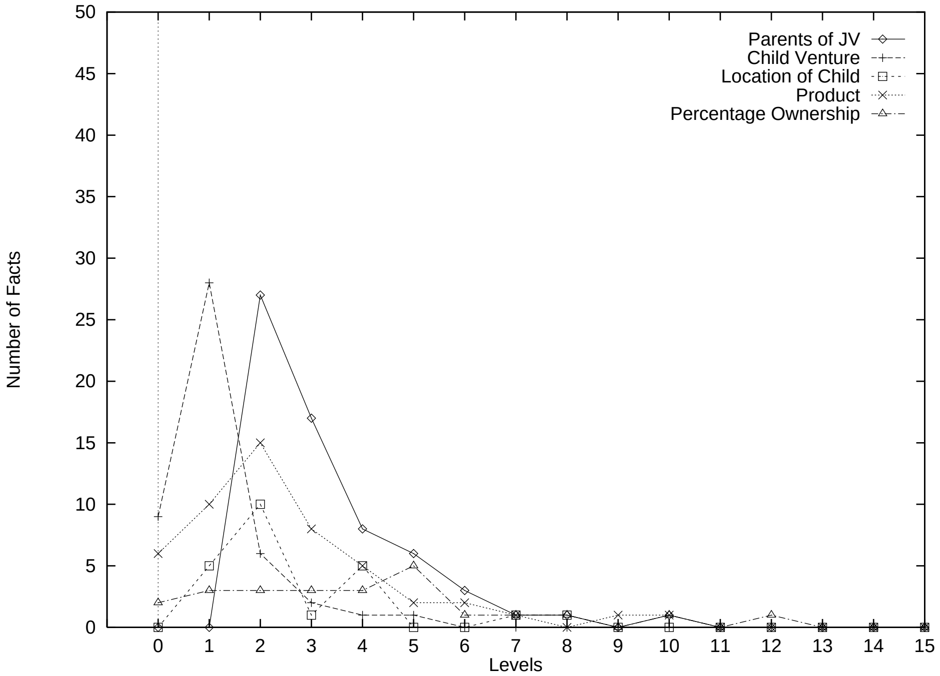

Figure 9: MUC-5: Level Distribution of Each of the Five Facts

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.

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.)

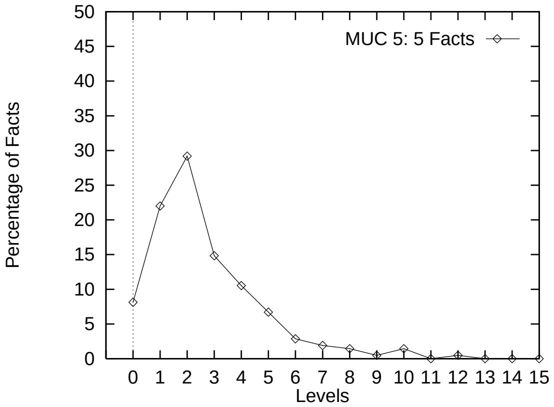

Figure 10: MUC-5: Level Distribution of the Five Facts Combined

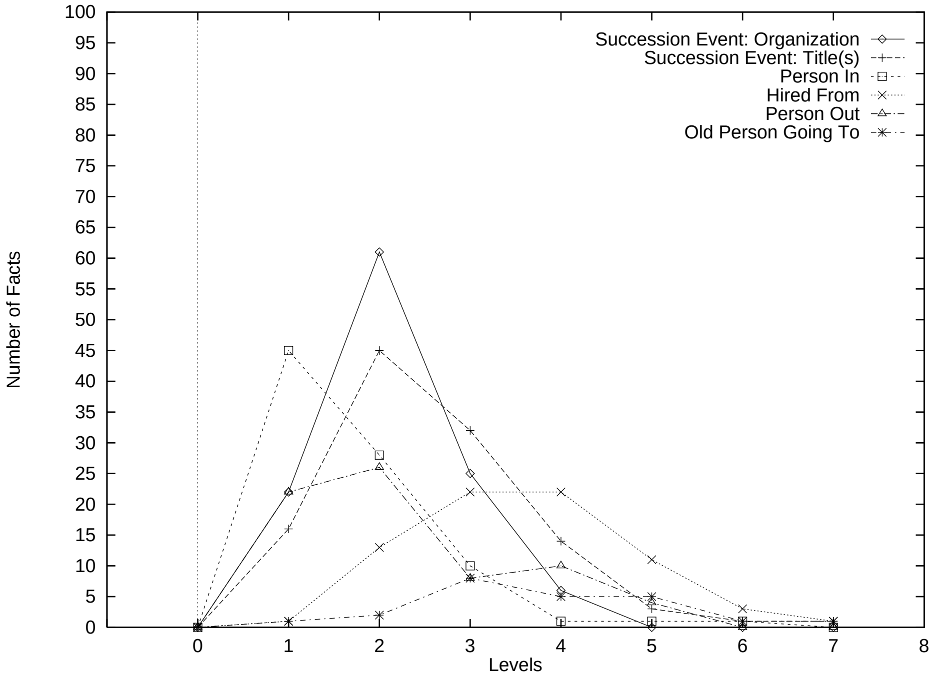

Figure 11: MUC-6: Level Distribution of Each of the Six Facts

- 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 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 9. The level distribution of the five facts combined is shown in Figure 10. Based on Figure 9 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 10 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 indicates that the MUC-5 EJV task was much harder than the MUC-4 task. In comparison, an analysis done by Beth Sundheim, using the features described earlier, 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].

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.)

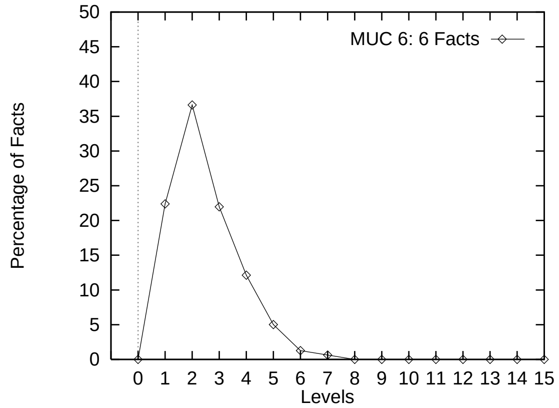

Figure 12: MUC-6: Level Distribution of the Six Facts Combined

MUC	Domain	Domain Numbers (in LUs)	Highest P&R F-Measure
MUC-4	Terrorist Attacks	1.87	55.93%
MUC-5	Joint Ventures	2.67	52.75%
MUC-6	Changes in Management Personnel	2.47	56.40%
MUC-7	Air Vehicle Launch Reports	2.44	

Figure 13: Domain Numbers of MUC-4, MUC-5, MUC-6, and MUC-7

- 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 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 11. The level distribution of these six facts combined is shown in Figure 12.

We calculated the domain number for the MUC-6 domain to be 2.47 LUs. Our analysis therefore indicates that the MUC-6 domain is almost as hard as the MUC-5 EJV domain.

Comparing the MUC Information Extraction Tasks

Figure 13 shows the domain numbers for the MUC tasks that have been analyzed. For each of the MUCs, the figure also shows the highest P&R F-Measure achieved by a system (for the information extraction task). Our analysis clearly separates the MUC-4 task from the later ones. The tasks for the later MUCs, however, have surprisingly similar complexity profiles: 20 to 30% percent level-1 facts, substantially higher level-2 facts, and decreasing values for higher level facts. Given that the analysis has been only done for 4 tasks, we do not want to infer too much from the difference between the complexities of the MUC-5 task and the MUC-6 and the MUC-7 tasks (which have roughly the same complexities).

Conclusion

The level of a fact with respect to a network for a piece of text provides a new method of classifying a fact based on the degree of syntactic complexity related to a fact. This classification tries to quantify the complexity of the syntax that holds the information in text. Moreover, studies based upon this classification scheme indicate that there is a correlation between the level of a fact and the difficulty of its extraction. 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.

This measure is just the beginning of a search for useful complexity measures. Despite its shortcomings, the current measure does quantify complexity on one very important dimension, namely the number of clauses (or phrases) required to specify a fact. For the short term it appears to be the best available vehicle for understanding the complexity of extracting a fact as shown in this paper.

Acknowledgments

The author would like to thank Ralph Grishman for the opportunity to analyze the MUC-7 information extraction task. The author also thanks Beth Sundheim for providing him with the official data from MUC-4, MUC-5, and MUC-6.

The author was supported by a Fellowship from IBM Corporation.

REFERENCES

[Bagga 1998] Bagga, Amit, and Alan W. Biermann. Analyzing the Performance of Message Understanding Systems, to appear in *Journal of Computational Linguistics and Chinese Language Processing*, 1998.

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

[Hirschman 1992] 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.

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

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

[Schubert 1979] 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 1993] 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 1995] 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.