Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 1-26

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.

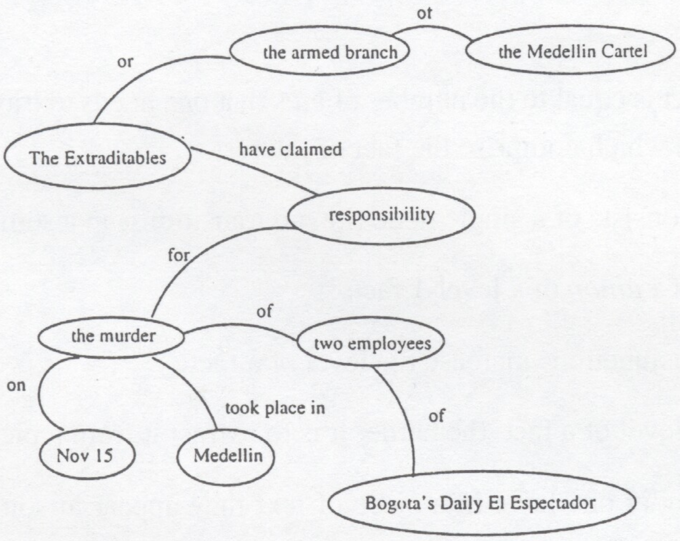


Figure 1 A Sample Semantic Network

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

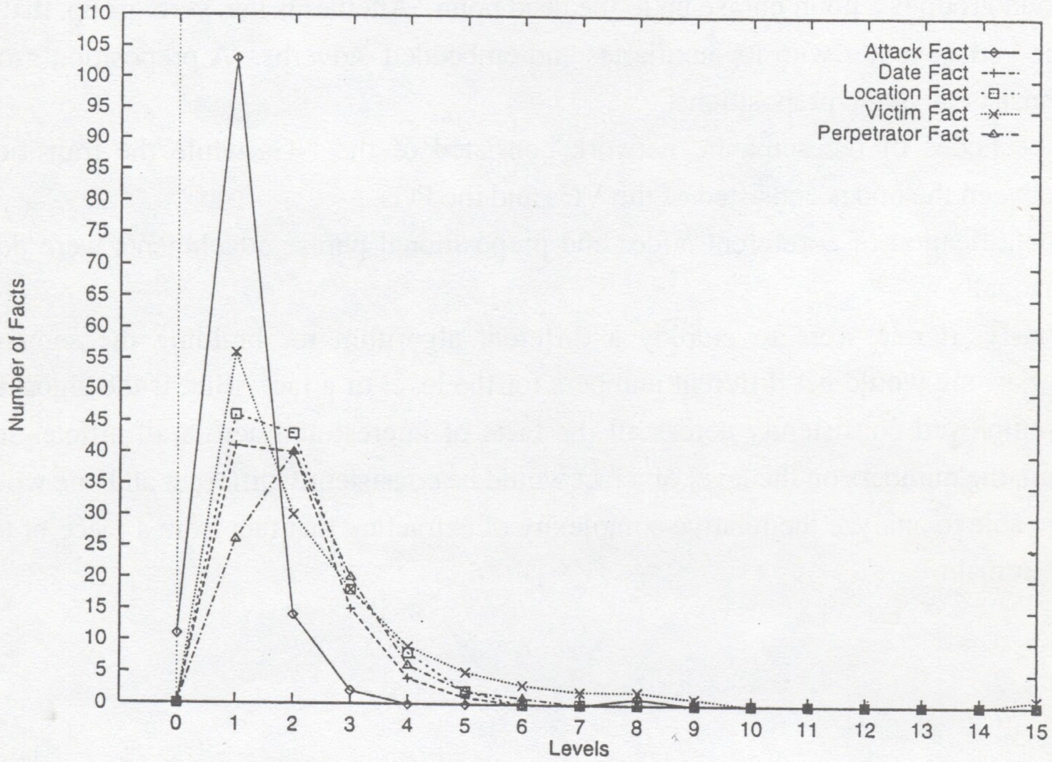


Figure 2 *MUC-4: Level Distribution of Each of the Five Facts*

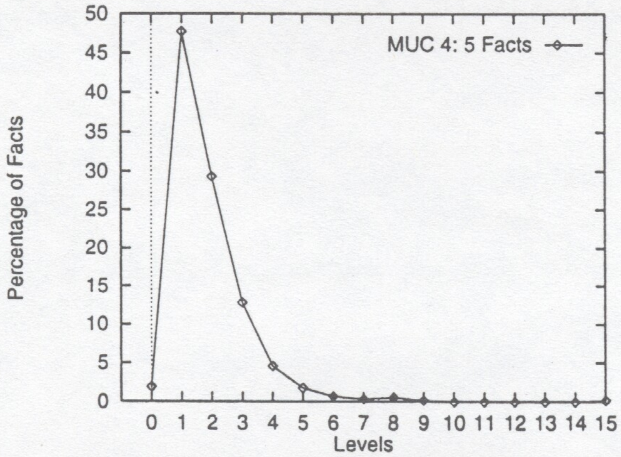


Figure 3 *MUC-4: Level Distribution of the Five Facts Combined*

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:

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

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

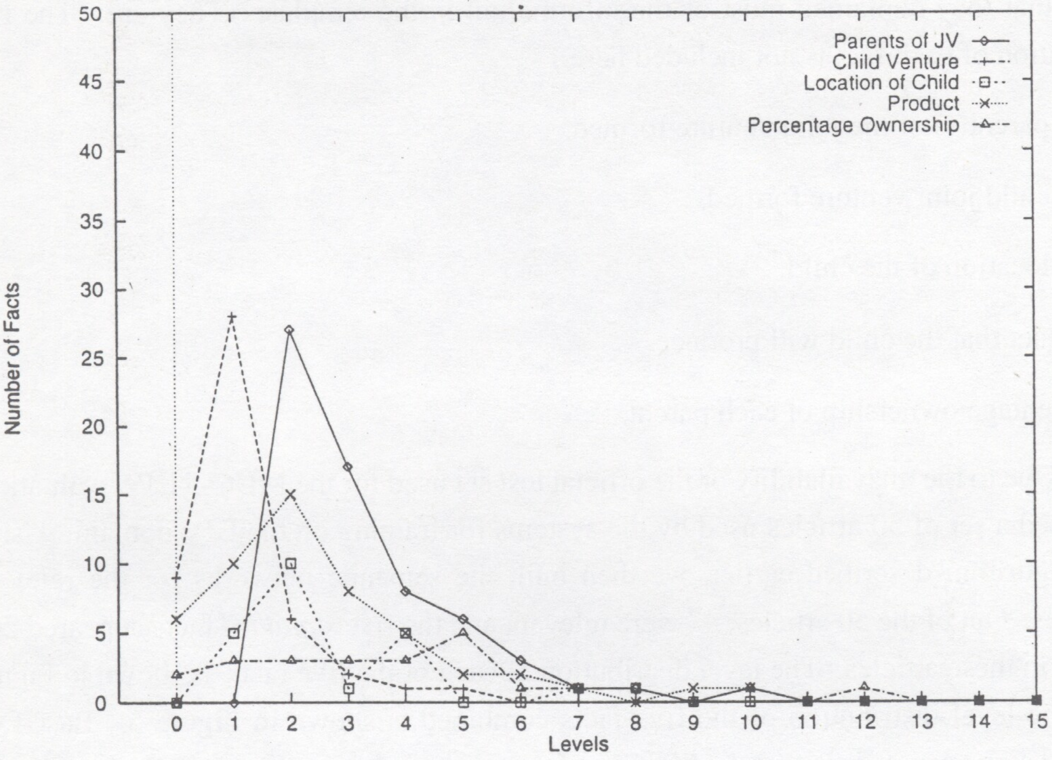


***Figure 4** MUC-5: Level Distribution of Each of the Five Facts*

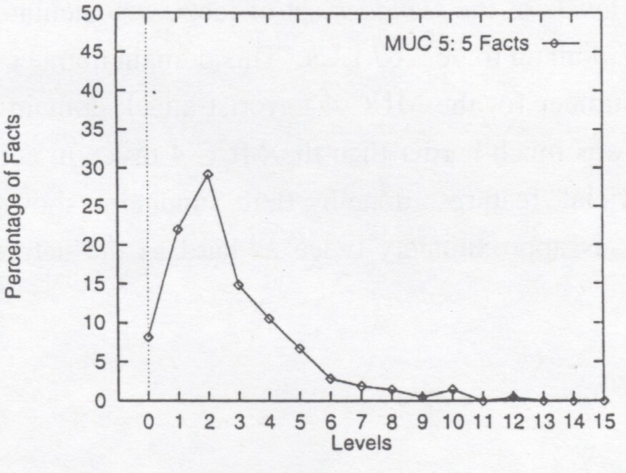


***Figure 5** MUC-5: Level Distribution of the Five Facts Combined*

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

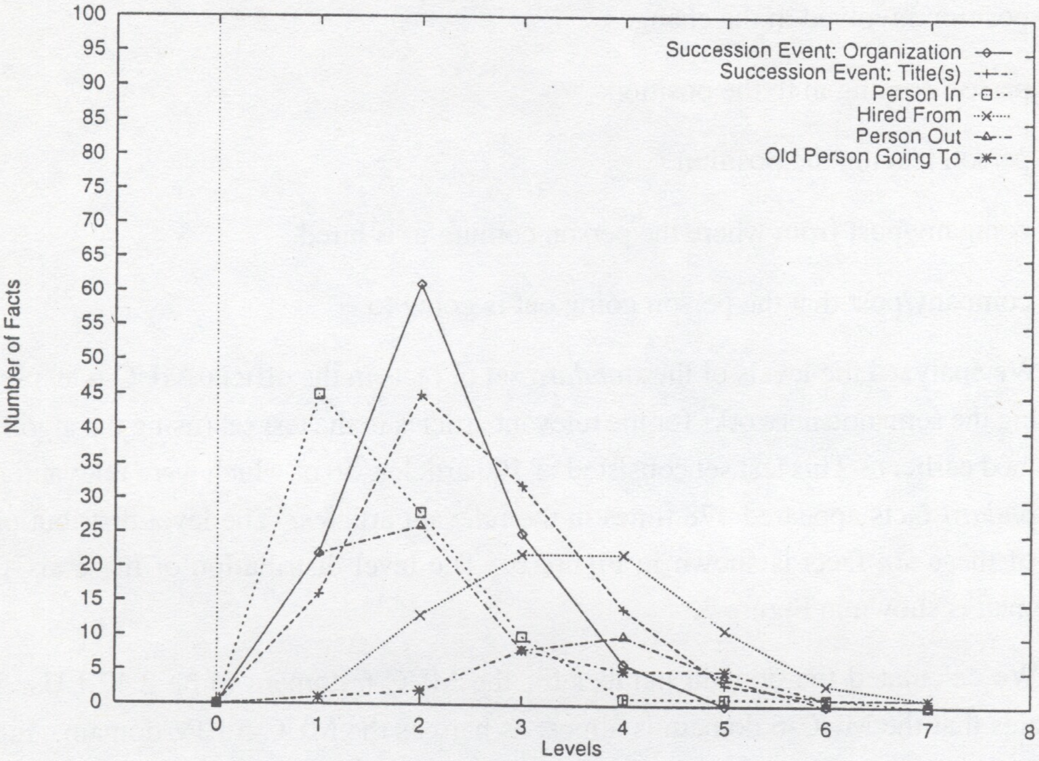


Figure 6 *MUC-6: Level Distribution of Each of the Six Facts*

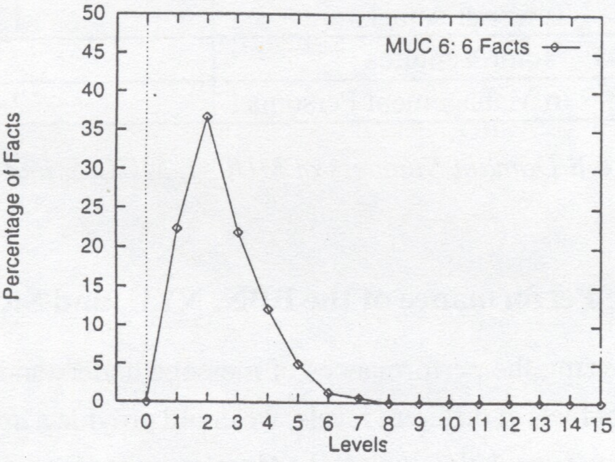


Figure 7 *MUC-6: Level Distribution of the Six Facts Combined*

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

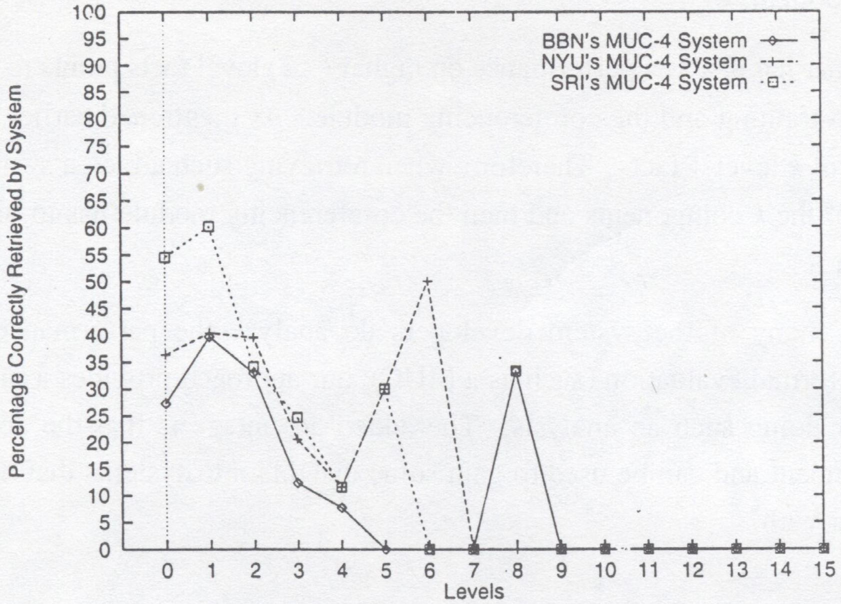


Figure 9 *Performance of the three systems at MUC-4*

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

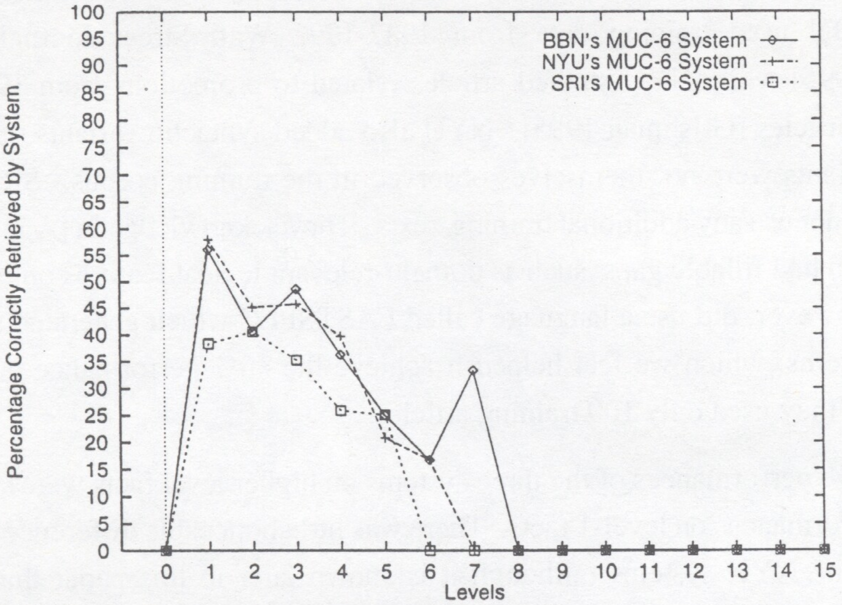


Figure 11 *Performance of the three systems at MUC-6*

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.

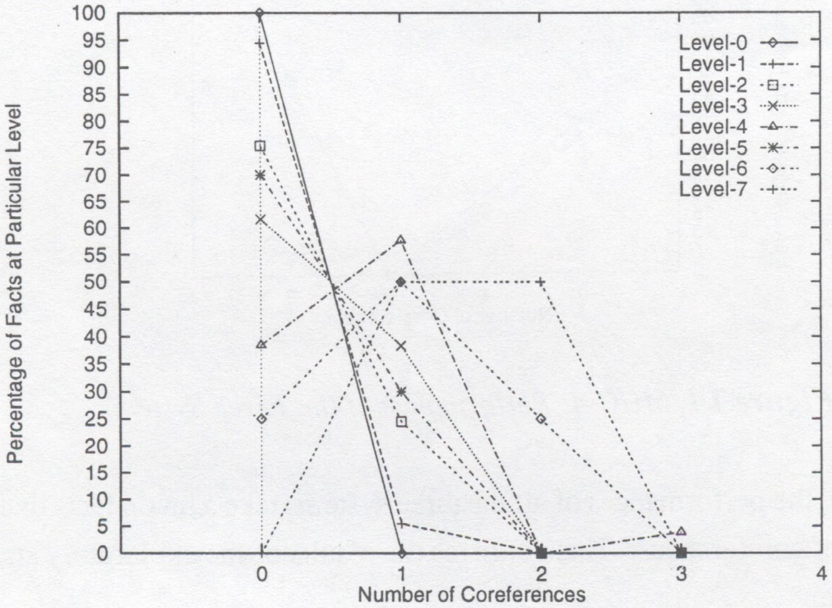


Figure 12 *MUC-4: Number of Coreferences At Each Level*

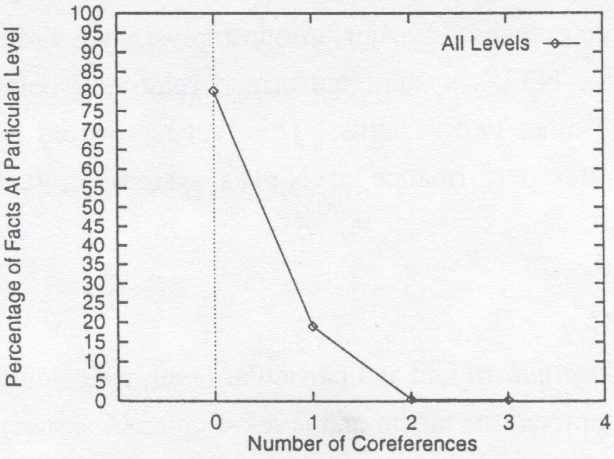


Figure 13 *MUC-4: Number of Coreferences At All Levels*

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.

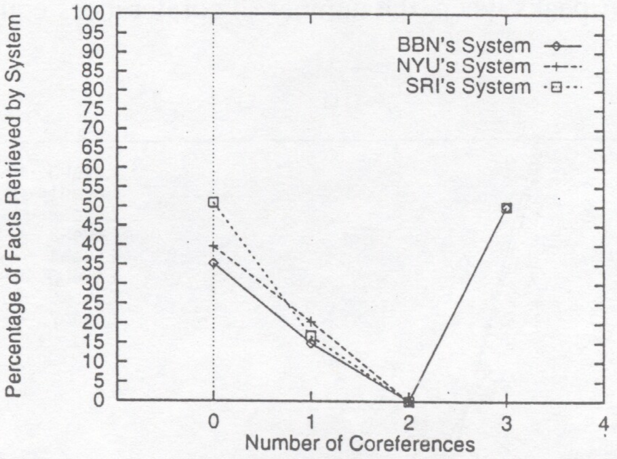


Figure 14 MUC-4: Performance of the Three System

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).

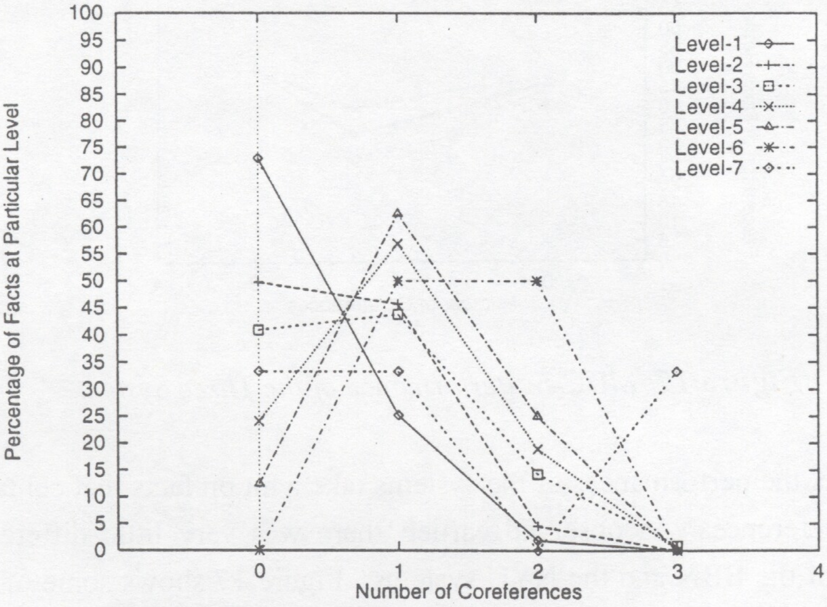


Figure 15 MUC-6: Number of Coreferences At Each Level

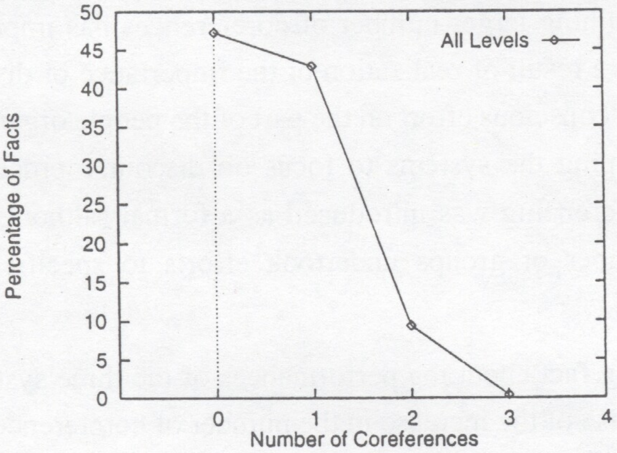


Figure 16 MUC-6: Number of Coreferences At All Levels

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.

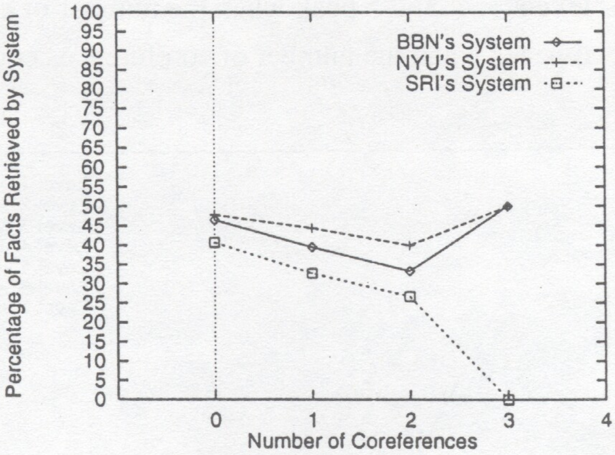


Figure 17 *MUC-6: Performance of the Three System*

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.

Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 27-44
©Computational Linguistics Society of R. O. C.

Unknown Word Detection for Chinese by a Corpus-based Learning Method

Keh-Jiann Chen[*], Ming-Hong Bai[*]

Abstract

One of the most prominent problems in computer processing of the Chinese language is identification of the words in a sentence. Since there are no blanks to mark word boundaries, identifying words is difficult because of segmentation ambiguities and occurrences of out-of-vocabulary words (i.e., unknown words). In this paper, a corpus-based learning method is proposed which derives sets of syntactic rules that are applied to distinguish monosyllabic words from monosyllabic morphemes which may be parts of unknown words or typographical errors. The corpus-based learning approach has the advantages of: 1. automatic rule learning, 2. automatic evaluation of the performance of each rule, and 3. balancing of recall and precision rates through dynamic rule set selection. The experimental results show that the rule set derived using the proposed method outperformed hand-crafted rules produced by human experts in detecting unknown words.

1. Introduction

One of the most prominent problems in computer processing of Chinese language is the identification of the words in a sentence. There are no blanks to mark word boundaries in Chinese text. As a result, identifying words is difficult because of segmentation ambiguities and occurrences of out-of-vocabulary words (i.e., unknown words). For instance, in (1), the proper name 王英雄 'Wang, Ying-Xiong' is a typical example of an unknown word, and it has ambiguous segmentation of 王 'king' 英雄 'hero'. Another example in (1) 台灣大學生 'university student in Taiwan' also has ambiguous segmentations of 台灣 'Taiwan' 大學生 'university student' , 台灣大學 'National Taiwan University' 生 'give birth to' ,and 台灣 'Taiwan' 大學 'university' 生 'give birth to' etc.:

(1) 王英雄是一個典型的台灣大學生。
'Ying-Xiong Wang is a typical university student in Taiwan.'

[*] Institute of Information Sicence, Academia Sinica, Taipei, Taiwan, R. O. C. E-mail: {kchen, evan}@iis.sinica.edu.tw

Most of the papers dealing with the problem of word segmentation have focused only on the resolution of ambiguous segmentation. The problem of unknown word identification is considered to be more difficult and needs to be further investigated. According to an inspection of the Sinica corpus [Chen et al., 1996], which is a balanced Chinese corpus with words segmented based on the Chinese word segmentation standard for information processing proposed by ROCLING [Huang et al., 1997], the most productive unknown words are of the following types.

1.1 Types of Unknown Words

Unknown words are defined as the words which are not in the lexicon. The following types of unknown words most frequently occur in the Sinica corpus. Table 1 shows the frequency distribution of unknown words of the most frequent 14 categories by examining 3 million-word data from the Sinica corpus.

(a) abbreviation (acronym): e.g., 中油 'China-fuel' (Nb) and 台汽 'Taiwan-bus' (Nb). (Please refer to table 1 for the meaning of each category name; for instance, Nb denotes proper names.)

It is difficult to identify abbreviations since their morphological structures are very irregular. Their affixes more or less reflect the conventions of the selection of meaning components [Huang 94]. However, the affixes of abbreviations are common words which are least informative for indicating the existence of unknown words.

(b) proper names: e.g., 陳壽 'Chen-So' (Nb), 香檳城 'Champaign-city' (Nc), and 微軟 'micro-soft' (Nb).

Proper names can be further classified into 3 sub-categories, i.e., names of people, names of place, and names of organizations. Certain key words are indicators for each different sub-category. For instance, there are about 100 common surnames which are prefix characters of Chinese personal names. The district names, such as 市 'city', 鄉 'country' etc., frequently occur as suffixes of the names of places. Identification of company names is as difficult as that of abbreviations since there is no restriction on the choice of morpheme components.

(c) derived words: e.g., 電腦化 'computer-ize' (Vh).

Derived words have affix morphemes which are strong indicators.

(d) compounds: e.g., 轉赴 'turn-go'(VCL), 獲允 'receive-permission' (VE), 搜尋法 'search-method' (Na), and 電腦桌 'computer-desk' (Na).

A compounds is a very productive type of unknown word. Nominal and verbal compounds are easily coined by combining two words/characters. Since there are more than 5000 commonly used Chinese characters, each with idiosyncratic syntactic behavior, it is hard to derive a set of morphological rules to generate the set of Chinese compounds. To identify Chinese compounds is, thus, also difficult.

(e) numeric type compounds: e.g., 1986 年 '1986-year' (Nd), 三千 'three-thousand', and 19 巷 '19-lane' (Nc).

The characteristic of numeric compounds is that they contain numbers as major components. For instances, dates, time, phone numbers, addresses, numbers, determiner-measure compounds etc. belong to this type. Since digital numbers are the major components of unknown words of this type and their morphological structures are more regular, they can be identified using the morphological rules.

Category	Frequency	Meaning of Category
A	1453	/*non-predictive adjective*/
Na	34372	/*common noun*/
Nb	14813	/*proper noun*/
Nc	9688	/*location noun*/
Nd	2264	/*time noun*/
VA	6466	/*active intransitive verb*/
VC	8462	/*active transitive verb*/
VCL	811	/*active transitive verb with locative object*/
VD	448	/*ditransitive verb*/
VE	1051	/*active transitive verb with sentential object*/
VG	996	/*classificatory verb*/
VH	10492	/*stative intransitive verb*/
VHC	498	/*stative causative verb*/
VJ	1471	/*stative transitive verb*/
total:	93,285	

Table 1. *The frequency distribution of unknown words in the most frequent categories.*

From the above discussion, it is seen that identification for each different type of unknown word is difficult and might require adopting different approaches. However, the processes for detecting the occurrences of each different type of unknown word are almost the same since they are all composed of morphemes of characters. In this paper, we focus only on the detection processes and leave the complete identification problem for future research.

1.2 Unknown Word Detection

Unknown words cause segmentation errors because out-of-vocabulary words in an input text normally are incorrectly segmented into pieces of single character word or shorter words. For instance, example (1) would be segmented into (2) after dictionary look-up and resolution of ambiguous segmentation:

(2)	王	英雄	是	一	個	典型	的	台灣	大學生。
	king	hero	be	DET	CL	typical	DE	Taiwan	university-student

It is difficult to know when an unknown word is encountered since all Chinese characters can either morphemes or words and there are no blanks to mark word boundaries. Therefore, without (or even with) syntactic or semantic checking, it is difficult to tell whether a character in a particular context is a part of an unknown word or whether it stands alone as a word. As mentioned in section 1.1, compound words and proper names are the two major types of unknown words. It is not possible to list all of the compounds in the lexicon nor possible to write simple rules which can enumerate the compounds without over-generation or under-generation. Each different type of compound must be identified using either content or context dependent rules. Proper names and their abbreviations have less content regularity. Identifying them relies more on contextual information. The occurrence of typographical errors makes the problem even more complicated. There is currently no satisfactory algorithm for identifying both unknown words and typographical errors, but researchers are separately working on each different type of problem.

Chang et al.[Chang et al., 94] used statistical methods to identify personal names in Chinese text and achieved a recall rate of 80% and a precision rate of 90%. Similar experiments were reported in [Sun et al., 94]. Their recall rate was 99.77% but with a lower precision of 70.06%. Both papers default with the recognition of Chinese personal names only. Chen & Lee [Chen & Lee 94] used morphological rules and contextual information to identify the names of organizations. Since organizational names are much more irregular than personal names in Chinese, they achieved a recall rate of 54.50% and a precision rate of 61.79%. A pilot study on automatic correction of Chinese spelling errors was done by Chang [Chang 94]. He used mutual information between a character and its neighboring words to detect spelling errors and to then automatically make the necessary corrections. The error detection process achieved a recall rate of 76.64% and a precision rate of 51.72%. Lin et al. [Lin et al., 93] did a preliminary study of the problem of unknown word identification. They used 17 morphological rules to recognize regular compounds and a statistical model to deal with irregular unknown words, such as proper names etc. With this unknown word resolution procedure, an error reduction rate of 78.34% was obtained for the word segmentation process. Since there is no standard reference data, the accuracy rates claimed in different papers vary due to different segmentation standards. In this study, we used the Sinica corpus as a standard reference data. As mentioned before, the Sinica corpus is a word-segmented corpus based on the Chinese word segmentation standard for information processing proposed by ROCLING. Therefore, it contains both known words and unknown words which are properly segmented, i.e., separated by blanks. The corpus was utilized for the purposes of training and testing. For unknown word and typographical error identification, the following two steps are proposed. The first step is to detect the existence of unknown words and typographical errors. The second step is the recognition process, which determines the type and boundaries of each unknown word and recognizes typographical errors. The reasons for separating the detection process from the recognition process are as follows:

a. For different types of unknown words and typographical errors, they may share the same detection process but have different recognition processes.

b. If the common method for spell checking is followed, an unknown word would be detected first, and a search for the best matching words performed next. Recognizing a Chinese word is somewhat different from spell checking, but they have a lot in common.

c.If the detection process performs well, the recognition process is better focused, making the total performance more efficient.

This paper focuses on the unknown word detection problem only. (Note that a typographical error is considered as a special kind of unknown word.) The unknown word detection problem and the dictionary-word detection problem are complementary problems since if all known words in an input text can be detected, then the rest of the character string will be unknown words. However, this is not a simple task since there are no blanks to delimit known words from unknown words. Therefore, the word segmentation process is applied first, and known words are delimited by blanks. Since unknown words are not listed in the dictionary, they will be segmented into shorter character/word sequences after a conventional dictionary-look-up word segmentation process. Sentence (3.b) shows the result of the word segmentation process on (3.a):

(3) a. 筑波大學延請七三年諾貝爾物理學獎得主江崎出任校長，

'The University of Tsukuba invited the winner of the '73 Nobel Award in Physics, Esaki, to be the Principal.'

b. 筑 波 大學 延請 七三年 諾貝爾 物理學 獎 得主
Tsuku -ba university invite '73 Nobel physics award winner
江 崎 出任 校長，
Esa -ki be principal.

According to an examination of a group of testing data which is a part of the Sinica corpus, 4572 occurrences out of 4632 unknowns were incorrectly segmented into sequences of shorter words, and each sequence contained at least one monosyllabic word. That is, 60 of the unknown words were segmented into sequences of multi-syllabic words only. Therefore, occurrences of monosyllabic words (i.e., single character words) in the segmented input text may denote the possible existence of unknown words. This is reasonable since it is very rare for compounds or proper names to be composed of several multi-syllabic words. Therefore, the process of detecting unknown words is equivalent to making a distinction between monosyllabic words and monosyllabic morphemes which are parts of unknown words. Hence, the complementary problem of unknown word detection is the problem of monosyllabic known-word detection. If all of the occurrences of monosyllabic words are considered as possible morphemes of unknown words, the precision of prediction is very low. When the word segmentation process was applied to the testing data taken from the Sinica corpus using a conventional dictionary look-up method, 69733 occurrences of monosyllabic words were found, but only 9343 were parts of unknown words, a precision of 13.40%. In order to improve the precision, monosyllabic words, which properly fit the contextual environment, should be identified and should not be considered as possible morphemes of unknown words. In the next section, the corpus-based learning approach to identification of contextually-proper monosyllabic words is introduced. In section 3, experimental results are presented, including a performance comparison between a hand-crafted method and the proposed corpus-based learning method.

2. Corpus-based Rule Learning for Identifying Monosyllabic Words

The procedure for detecting unknown words is roughly divided into three steps: 1. word segmentation, 2. part-of-speech tagging, and 3. identification of contextually-proper monosyllabic words. The word segmentation procedure identifies words using a dictionary look-up method and resolves segmentation ambiguities by maximizing the probability of a segmented word sequence [Chiang 92, Chang 91, Sproat 96] or by using heuristic methods [Chen 92, Lee 91]. Either method can achieve very satisfactory results.

Both have an accuracy rate of over 99%. For the purpose of unknown word identification, some regular types of compounds, such as numbers, determinant-measure compounds, and reduplication, which have regular morphological structures, are also identified by means of their respective morphological rules during the word segmentation process [Chen 92, Lin 93]. The second step, part-of-speech (pos) tagging, is carried out to support step3 and the later process of unknown word identification. After pos tagging, sentence (3.b) becomes sentence (4); each word contains a unique pos:

(4) 筑 (BOUND) 波 (Nf) 大學 (Nb) 延請 (VC) 七三年 (DM) 諾貝爾 (Nb)
Tsuku -ba university invite '73 Nobel
物理學 (Na) 獎 (Na) 得主 (Na) 江 (Na) 崎 (BOUND) 出任 (VG)
physics award winnter Esa -ki be
校長 (Na) ，
principal.

Although the pos sequence may not be 100% correct, it is the most probable pos sequence in terms of pos bi-gram statistics [Liu 95]. The details of the first two steps are not the major concern of this paper. The focus here is on the step of identifying contextually-proper monosyllabic words. Hereafter, for simplicity, the term 'proper-character' will denote a contextually-proper monosyllabic word, and the term 'improper-character' will be used to denote a contextually-improper monosyllabic word which might be part of an unknown word. The way to identify proper-characters is to check the following properties:

(1) a proper-character should not be a bound-morpheme, and
(2) the context of a proper-character should be grammatical.

Hence, if a character is a bound-morpheme, it will be considered as possibly being an unknown word. However, almost any Chinese character can function either as a word or as a bound morpheme. A character's functional role is contextually dependent. Therefore, every monosyllabic word should be checked in its context for grammaticality by means of syntactic or semantic rules. For processing efficiency, such rules should be simple and have only local dependencies. It is not feasible to parse whole sentences in order to check whether or not characters are proper-characters. The task is then to derive a set of rules which can be used to check the grammaticality of characters in context. If the rules are too stringent, then too many proper-characters will be considered as improper-characters, resulting in a low precision rate. On the other hand, if the rules are too relaxed, then too many improper-characters will be considered as proper-characters, resulting in a low recall rate. Therefore, there is a tradeoff between recall and precision. In the case of unknown word detection, a higher recall rate and an acceptable precision rate is preferred. Writing handcrafted rules is difficult because there are more than 5000 commonly used Chinese characters, and each of them may behave differently. A corpus-based learning approach is adapted to derive the set of contextual rules and to select the best set of rules by evaluating the performance of each individual rule. The approach is very similar to the error-driven learning method proposed by Brill [Brill 95]. Before the learning method is introduced, two commonly used measures for unknown word detection are defined. These two performance measures will be used throughout the paper:

Recall Rate = # of unknown word detected / total number of unknowns;

Precision Rate = # of correctly detected improper-characters / total # of guesses.

There are two types of unknown words. Type one unknown words include monosyllabic morphemes. Type two unknown words are composed with multi-syllabic words only. Only the detection of type one unknown words is considered here since type two unknown words occur very rarely as we mentioned before. An unknown word is considered successfully detected if any one of its components is detected as an improper-character. It is noted that the numerators for the recall rate and the precision rate are different since if two (or more) components of an unknown word are detected as improper-characters, it is reasonable to count only one word detection but two improper-character detections. For the corpus-based learning method, a training corpus with all the words segmented and pos-tagged is used. The monosyllabic words in the training corpus are instances of proper-characters, and the words in the training corpus which are not in the dictionary are instances of unknown words. Segmenting the unknown words using a dictionary look-up method produces instances of improper-characters. By examining the instances of proper and improper characters and their contexts, the rule patterns and their performance evaluations can be derived and can be represented as triplets (rule pattern, total # of matched instances, # of improper instances). Examples are shown in Appendix1. A contextual dependent rule may be: a uni-gram pattern, such as '{ 的 }', '{ 好 }', '{(Nh)}', '{(T)}', a bi-gram patterns, such as '{ 會 } 覺得 ', '{ 就 }(VH)', '(Na){ 上 }', '{(Dfa)}(Vh)', '(Ve){(Vj)}', or a tri-gram patterns, such as '{ 極 }(VH)(T)', '(Na)(Dfa){ 高 }', where the string in the curly brackets will match a proper-character and the other parts will match its context.

A good rule pattern has high applicability and high discrimination value (i.e., it occurs frequently and matches either proper-characters or improper-characters only, but not both). In fact, no rule has perfect discriminating ability. For instance, the rule '(Na){(Nb)}' can be applied to ' 會計 (Na) 劉 (Nb)' in (5) and ' 院士 (Na) 閻 (Nb)' in (6). The results are correct in (5) and incorrect in (6):

(5)	會計 (Na)	人員 (Na)	劉 (Nb)	小姐 (Na)
	accounting	staff	Liu	Miss.

(6)	院士 (Na)	閻 (Nb)	振興 (VC)	先生 (Na)
	academician	Yan	-strengthen	Mr.

Therefore, a greedy method is adopted in selecting the best set of unknown word detection rules. A set of rules which can identify proper-characters with high accuracy is selected. The rules with applicability greater than a threshold value are sequentially chosen according to the order of their accuracy. The rules for identifying improper-characters was not used because most improper-characters are of low frequency. Conversely, the selected rules were used as the recognition rules for proper-characters. A character matched by any one of the selected rules is considered a proper-character. Characters which are not matched by any one of the rules are considered candidates of improper-characters.

Rule selection algorithm:

1. Determine the threshold values for rule accuracy and applicability. For each rule Ri, when applied on the training corpus, the rule accuracy(Ri) = Mi / Ti, where Mi is the # of instances of matches of Ri with proper characters; Ti is the total # of matches of Ri. The rule applicability(Ri) = Ti.
2. Sequentially select the rules with the highest rule accuracy and applicability greater than the threshold value until there are no rules satisfying both threshold values.

The threshold value for rule accuracy controls the precision and recall performance of the final selected rule set. A higher accuracy requirement means fewer improper-characters will be wrongly recognized as proper-characters. Therefore, the performance of such a rule set will have a higher recall value. However, those proper-characters not matched with any rules will be mistaken as improper-characters, which lowers precision. On the other hand, if a lower accuracy threshold value is used, then most of the proper-characters will be recognized, and many of the improper-characters will also be mistakenly recognized as proper-characters, resulting in a lower recall rate and possibly a higher precision rate before reaching the maximal precision value. Therefore, if a detection rule set with a high recall rate is desired, the threshold value of rule accuracy must be set high. If precision is more important, then the threshold value must be properly lowered to an optimal point. A balance between recall and precision should be considered.

In the next section, the experimental results of different threshold values are presented. The threshold value for rule applicability controls the number of rules to be selected and ensures that only useful rules are selected.

The selected rule type may subsume another. Shorter rule patterns are usually more general than longer ones. There are redundant rules in the initial rule selection. A further screening process is needed to remove the redundant rules. The screening process is based on the following fact: if a rule Ri is subsumed by rule Rj, then pattern of Ri is a sub-string of pattern Rj. For example the rule '{ 的 }' is more general than the rule '{ 的 } (Na)'. If the rule '{ 的 }' is selected, then the rule '{ 的 } (Na)' is redundant and can be removed from the rule set. Since a character matched by any one of the selected rules is considered a proper-character, more specific rules will be redundant and only the most general rules will remain after the screening process.

Screening Algorithm:

1. Sort the rules according to their string patterns in increasing order, resulting in rules R1...Rn.
2. For i from 1 to n, if there is a j such that j< i, and Rj is a sub-string of Ri, then remove Ri.

3. Experimental Results

The corpus-based learning method for unknown word detection was tested on the Sinica corpus. The Sinica corpus version 2.0 contains 3.5 million words. 3 million words were used as the training corpus and 0.15 million words for the testing corpus. The word entries in the CKIP lexicon were considered as known words. The CKIP lexicon contains about 80,000 entries of Chinese words with their syntactic categories and grammatical information [CKIP 93]. A word is considered as an unknown word if it is not in the CKIP lexicon and is not identified by the word segmentation program as a foreign word (for instance, English), a number, or a reduplicated compound. There were 93285 unknown words in the training corpus and 4632 unknown words in the testing corpus. A few bi-word compounds were deliberately ignored as unknowns, such as 分析化學 'analytical chemistry' and 技術人員 'technical member', since they are not identifiable by any algorithm which does not incorporate real world knowledge. In addition, whether these are single compounds or noun phrases made up of two words is debatable. In fact, ignoring bi-word compounds did not affect the results very much since the fact that there were only 60 such unknown words out of 4632 shows that they rarely occurred in the corpus.

The following types of rule patterns were generated from the training corpus. Each rule contains a token within curly brackets and its contextual tokens without brackets. For some rules, there may be no contextual dependencies.

Rule type	Examples
char	{的}
word char	不 {願}
char word	{全} 世界
category	{(T)}
{category} category	{(Dfa)} (Vh)
category {category}	(Na) {(Vcl)}
char category	{就} (VH)
category char	(Na) {上}
category category char	(Na) (Dfa) {高}
char category category	{極} (Vh) (T)

Rules of the 10 different types of patterns above were generated automatically by extracting each instance of monosyllabic words in the training corpus. Every generated rule pattern was checked for redundancy, and the frequencies of proper and improper occurrences were tallied. For instance, the pattern '{ 的 }' occurred 165980 times in the training corpus; 165916 of these were proper instances and 64 of these were improper instances (i.e., ' 的 ' occurred 64 times as part of an unknown word). Appendix 1 shows some of the rule patterns and their total occurrences counts as well as the number of improper instances. In the initial stage, 1455633 rules were found. After eliminating rules with frequency less than 3, 215817 rules remained. In the next stage, different rule selection threshold values were used to generate 10 different sets of rules. These rule sets were used to detect unknown words in the testing corpus. The testing corpus contained 152560 words. In the first step, the running text of the testing corpus was segmented into words using a dictionary look-up method which were then tagged with their part-of-speech by an automatic tagging process. Each different rule set was applied to detect the unknown words in the testing corpus. A character without a match was considered as part of an unknown word. Appendix 2 shows some examples. The performance results of different rule sets are shown in Table 2, and the detail statistics are shown in Appendix 3.

Rule selection criteria	Recall rate	Precision rate	# of rules after screening
(0) no rule applied	100%	13.40%	0
(1) rule accuracy >= 55%	63.32%	73.69%	3054
(2) rule accuracy >= 60%	63.89%	73.73%	3239
(3) rule accuracy >= 65%	64.85%	74.04%	5209
(4) rule accuracy >= 70%	68.18%	74.61%	6081
(5) rule accuracy >= 75%	73.80%	74.36%	8611
(6) rule accuracy >= 80%	77.34%	73.26%	10500
(7) rule accuracy >= 85%	81.06%	71.52%	13962
(8) rule accuracy >= 90%	87.40%	68.74%	18967
(9) rule accuracy >= 95%	93.66%	64.73%	31309
(10) rule accuracy >= 98%	96.30%	60.62%	45839

Note: all of the applicability values are set to rule frequency >= 3.

Table 2. *The experimental results of unknown word detection on the testing corpus.*

The results show that there is a tradeoff between precision and recall rate, but that the overall performance was much better than that of the handcraft rules written by human experts. They examined the training corpus and wrote up a rule set for proper-characters to the best of their ability. The handcraft rules had a precision rate of 39.11% and a recall rate of 81.45%, which are much lower than the rule set, made using the corpus-based rule learning method. The syntactic complexity of monosyllabic words was the reason for the lower coverage of the handcraft rules. Some handcraft rules are shown in Appendix 4. It is clearly shown that the handcraft rules suffer from low accuracy because a limited number of rules can be derived and the rules are usually too general to achieve high precision rates. There were only 139 hand-crafted rules while the proposed method generated thousands of rules as shown in Table 2. The number of rules selected increased with the decrement of the accuracy of the rule selection criteria because more rules satisfied the lower accuracy requirement. However, the number of rules after the screening process was lower in accordance with the decrement of the accuracy of the rule selection criteria. For instance, 207059 and 210552 rules were selected, respectively, for the rule accuracy criterion of 98% and 95%, but after the screening process, the number of rules was 45839 and 31309. The reason for this is that achieving a higher accuracy requires more contextual dependency rules to discriminate between proper-characters and improper-characters. On the other hand, a lower accuracy requirement may cause the inclusion of many more short rules. This causes a lot of long rules to be subsumed by shorter rules eliminated during the screening process.

4. Conclusion and Future Research

The corpus-based learning approach proved to be an effective and easy method for finding unknown word detection rules. The advantages of using a corpus-based method are as follows:

a. The syntactic patterns of proper-characters are complicated and numerous. It is hard to hand-code each different pattern, yet most high frequency patterns are extractable from the corpus.

b. The corpus provides standard reference data not only for rule generation, but also for rule evaluation. The hand-craft rules can also be evaluated automatically and incorporated into the final detection rule set if the rule has a high accuracy rate.

c. It is easy to control the balance between the precision and the recall of the detection algorithm since we know the performance of each detection rule based on the training corpus.

Different types of unknown words have different levels of difficulty in identification. The detection of compounds is the most difficult because some of their morphological structures are similar to common syntactic structures. The detection of proper names and typographical errors is believed to be easier because of their irregular syntactic patterns. The results with respect to different types of syntactic categories were checked. Appendix 3 shows that the recall rates of proper names (i.e., category Nb) were less affected by the higher precision requirement. There was no data for typographical errors, but the detection of typographical errors is believed to be similar to the detection of proper names; that is, a higher precision can be achieved without sacrificing the recall rate. If parallel corpora with and without typographical errors are available, the corpus-based rule learning method can also be applied to the detection of typographical errors in Chinese.

After the unknown word detection process, an identification algorithm will be required to find the exact boundaries and the part-of-speech of each unknown word. This will require future research. Different types of rules will be required in identifying different compounds and proper names. The corpus can still play an essential role in the generation of rules and in their evaluation.

Acknowledgments

The authors wish to thank Dr. Charles Lee and the anonymous reviewers for their useful comments on this paper.

References

Brill, Eric, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics* 21.4 (1995), pp.543-566.

Chang, J. S., C. D. Chen, & S. D. Chen, "Word Segmentation through Constraint Satisfaction and Statistical Optimization," *Proceedings of ROCLING IV* (1991), pp. 147-165.

Chang, C. H., "A Pilot Study on Automatic Chinese Spelling Error Correction" *Communication of COLIPS*, 4.2 (1994), pp.143-149.

Chang J. S.,S.D. Chen, S. J. Ker, Y. Chen, & J. Liu,1994 "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts", *Computer Processing of Chinese and Oriental Languages*, 8.1(1994), pp.75-85.

Chen, H.H., & J.C. Lee, "The Identification of Organization Names in Chinese Texts", *Communication of COLIPS*, 4.2(1994), pp. 131-142.

Chen, K.J., C.R. Huang, L. P. Chang & H.L. Hsu, "SINICA CORPUS: Design Methodology for Balanced Corpora," *Proceedings of PACLIC 11th Conference* (1996), pp.167-176.

Chen, K.J. & S.H. Liu, "Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th Coling* (1992), pp. 101-107.

Chiang, T. H., M. Y. Lin, & K. Y. Su, "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING V* (1992), pp. 121-146.

Huang, C. R. Et al.,"The Introduction of Sinica Corpus," *Proceedings of ROCLING VIII* (1995), pp. 81-89.

Huang, C.R., K.J. Chen, & Li-Li Chang, "Segmentation Standard for Chinese Natural Language Processing," *International Journal of Computational Linguistics and Chinese Language Processing* 2.2 (1997), pp. 74-62.

Lee,H.J. & C.L. Yeh, "Rule-based Word Identification for Mandarin Chinese Sentences- A Unification Approach," *Computer Processing of Chinese and Oriental Languages*, 5.1 (1991), pp. 97-118.

Lin, M. Y., T. H. Chiang, & K. Y. Su, "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," *Proceedings of ROCLING VI* (1993), pp. 119-137.

Liu S. H., K. J. Chen, L.P. Chang, & Y.H. Chin, "Automatic Part-of-Speech Tagging for Chinese Corpora," *Computer Processing of Chinese and Oriental Languages*, 9.1(1995), pp.31-48.

Sproat, R., C. Shih, W. Gale, & N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22.3 (1996), pp.377-404.

Sun, M. S., C.N. Huang, H.Y. Gao, & Jie Fang, "Identifying Chinese Names in Unrestricted Texts", *Communication of COLIPS*, 4.2 (1994), pp. 113-122.

Appendix 1. Samples of rule patterns.

rule	frequency	error	accuracy
{的}	165980	64	99.71 %
{是}	41089	78	98.10 %
{也}	16066	11	99.31 %
{她}	6185	4	99.35 %
{這}	5046	1	99.80 %
{或}	4582	3	99.34 %
{該}	2302	2	99.13 %
{(T)}	177641	177	99.00 %
{(Nh)}	73034	344	99.53 %
{(Caa)}	46659	392	99.16 %
{(SHI)}	41089	78	99.81%
{(Dfa)}(VH)	11037	39	99.65 %
{(Nh)}(Na)	6640	62	99.07 %
{(P)}(Nh)	6247	52	99.17 %
{(Nep)}(Na)	4030	26	99.35 %
(Na){(VCL)}	8062	299	96.30 %
(VC){(Di)}	4155	76	98.18 %
(VE){(VJ)}	1884	46	97.56 %
(VJ){(VJ)}	1489	53	96.44 %
(VJ){(Dfa)}	1004	5	99.50 %
{與}(Na)	3933	6	99.85 %
{及}(Na)	2831	18	99.36 %
{在}(VC)	2451	2	99.92 %
(VH){地}	1787	14	99.22 %
(VC){者}	1731	1	99.94 %
(Na){很}	1172	0	100 %
{再}(VC)(Na)	221	0	100 %
{令}(Na)(VH)	200	0	100 %
{各}(Na)(Na)	190	3	98.42 %
{極}(VH)(T)	187	1	99.47 %
(Na)(Dfa){高}	263	0	100 %
(Na)(VH){地}	248	1	99.60 %
(Na)(Na){時}	231	2	99.14 %
(T)(Na){則}	174	0	100 %
{會}覺得	139	1	99.28 %
{才}知道	124	0	100 %
{拿}著	121	0	100 %
{迄}今	117	0	100 %
的{話}	1406	2	99.86 %
並{非}	319	0	100 %

Appendix 2. Samples of testing results.

The first line contains the original text. The second line shows the result of word segmentation and pos tagging. The third line is the result of unknown word detection, where improper-characters are marked with '(?)'.

```
*************************************
有的時候我想吃點美國菜，
有的 (Nepa)  時候 (Na)  我 (Nh)  想 (VE)  吃 (V)  點 (Na)  美國 (Nc)  菜 (Na),
有的 ()(Nepa)  時候 ()(Na)  我 ()(Nh)  想 ()(VE)  吃 ()(V)  點 ()(Na)  美國 ()(Nc)  菜 (?)(Na) ,
*************************************
微軟過去兩年也推出了近百種新產品，
微 (D) 軟 (VH)  過去 (Nd) 兩年 (DM) 也 (D)  推出 (VC)  了 (VJ)  近百種 (DM) 新 (VH) 產品 (Na),
微 ()(D)  軟 (?)(VH)  過去 ()(Nd)  兩年 ()(DM)  也 ()(D) 推出  ()(VC)  了 ()(VJ)  近百種  ()(DM)  新
()(VH) 產品 ()(Na),
********************************
即使營收和獲利成長開始減慢，
即使 (Cbb) 營收 (Na)  和 (Caa)  獲利 (VH)  成長 (VH)  開始 (VL)  減 (VJ)  慢 (VH) ，
即使 ()(Cbb) 營收 ()(Na)  和 ()(Caa) 獲利 ()(VH)  成長 ()(VH)  開始 ()(VL)  減 (?)(VJ)  慢 ()(VH) ，
*******************************
一九九四將是日本教育的改革年，
一九九四 (DM) 將 (D) 是 (SHI) 日本 (Nc) 教育 (VC) 的 (T) 改革 (VC) 年 (Nf) ，
一九九四 ()(DM) 將 ()(D) 是 ()(SHI) 日本 ()(Nc) 教育 ()(VC) 的 ()(T) 改革 ()(VC) 年 (?)(Nf) ，
*******************************
日本可能出現第一個個人主義世代。
日本 (Nc) 可能 (D) 出現 (VH) 第一個 (DM) 個 (Nf) 人 (Na) 主義 (Na) 世代 (Na) 。
日本 ()(Nc) 可能 ()(D) 出現 ()(VH) 第一個 ()(DM) 個 (?)(Nf) 人 (?)(Na) 主義 ()(Na) 世代 ()(Na) 。
*******************************
筑波大學延請七三年諾貝爾物理學獎得主江崎出任校長，
筑 (BOUND)  波 (Nf)  大學 (Nb)  延請 (VC)  七三年 (DM)  諾貝爾 (Nb)  物理學 (Na)  獎 (Na)  得
主 (Na)  江 (Na) 崎 (BOUND) 出任 (VG) 校長 (Na) ，
筑 (?)(BOUND)  波 (?)(Nf)  大學 ()(Nb)  延請 ()(VC)  七三年 ()(DM)  諾貝爾 ()(Nb)  物理學 ()(Na)  獎
(?)(Na)  得主 ()(Na)  江 (?)(Na) 崎 (?)(BOUND) 出任 ()(VG) 校長 ()(Na) ，
*******************************
就連整個體系中最官僚的教育當局──日本文部省，
就 (Da)  連  (D)  整個 (DM)  體系 (Na)  中 (Ng)  最 (Dfa)  官僚 (Na)  的 (T)  教育 (VC)  當局 (Na) —
(BOUND) 日本 (Nc) 文 (BOUND) 部 (Nc) 省 (Nc) ，
就 ()(Da) 連  ()(D) 整個  ()(DM) 體系  ()(Na) 中  ()(Ng)  最 ()(Dfa)  官僚 ()(Na)  的 ()(T)  教育 ()(VC)
當局 ()(Na) — ()(BOUND)  日本 ()(Nc)  文 (?)(BOUND)  部 (?)(Nc)  省 (?)(Nc) ，
*********************************
也在調整一向溫吞的改革步伐。
也 (D)  在 (VCL)  調整 (VC)  一向 (D)  溫 (VHC)  吞 (VC)  的 (T)  改革 (VC)  步伐 (Na) 。
也 ()(D)  在 ()(VCL)  調整 ()(VC)  一向 ()(D)  溫 (?)(VHC)  吞 (?)(VC)  的 ()(T)  改革 ()(VC)  步
伐 ()(Na) 。
***********************************
業者可以更準確地捕捉各個特定人口群，
業者 (Na)  可以 (D)  更 (D)  準確 (VH)  地 (Na)  捕捉 (VC)  各個 (DM)  特定 (A) 人口 (Na)  群
(Nf) ，
業者 ()(Na)  可以 ()(D)  更 ()(D)  準確 ()(VH)  地 ()(Na)  捕捉 ()(VC)  各個 ()(DM)  特定 ()(A) 人
口 ()(Na)  群 (?)(Nf) ，
***********************************
```

Appendix 3. The detailed performance results for the different rule sets.

The first column shows the categories of unknown words.

The second column is the number of occurrences of the unknown words in the category shown in column one.

The third column is the recall rates of the unknown words detected under different rule sets.

Category	# of Unknown Words	Frequency > 2						
		Accuracy >=						
		55%	60%	70%	80%	90%	95%	98%
A	63	66.67%	66.67%	66.67%	74.60%	79.37%	87.30%	96.83%
Na	1396	75.07%	76.29%	79.87%	85.24%	92.12%	95.85%	97.13%
Nb	1511	87.16%	87.56%	90.47%	95.90%	98.28%	99.47%	99.60%
Nc	424	67.92%	67.92%	74.76%	75.94%	89.86%	91.04%	95.52%
Nd	24	16.67%	16.67%	25.00%	37.50%	50.00%	79.17%	83.33%
Nh	62	4.84%	4.89%	35.48%	75.81%	88.71%	90.32%	93.55%
VA	151	31.79%	32.45%	34.44%	54.30%	69.54%	83.44%	86.76%
VB	25	20.00%	20.00%	24.00%	40.00%	64.00%	84.00%	84.00%
VC	439	14.58%	14.58%	20.05%	41.91%	73.13%	89.29%	94.99%
VCL	63	14.29%	14.29%	15.87%	36.51%	79.37%	90.48%	96.83%
VD	48	2.08%	2.08%	8.33%	56.25%	77.08%	89.58%	93.75%
VE	70	4.29%	4.29%	4.29%	12.86%	24.29%	78.57%	88.57%
VG	69	7.25%	7.25%	10.15%	21.74%	40.58%	69.57%	86.96%
VH	137	22.65%	24.09%	35.77%	60.58%	73.72%	84.67%	89.78%
VHC	23	91.30%	91.30%	91.30%	95.65%	95.65%	95.65%	95.65%
VJ	67	8.96%	8.96%	11.94%	25.37%	44.78%	67.16%	83.58%
Total:	4572							
Recall:		63.32%	63.89%	68.18%	77.34%	87.40%	93.66%	96.30%
Precision:		73.70%	73.73%	74.61%	73.27%	68.74%	64.73%	60.63%

Appendix 4. Some examples of the handcraft rules.

The items in the curly brackets match a proper-character and the items in the round brackets match its context according to their linear order. The symbol ',' in the rules denotes an 'or' relation.

1. (之 , 的){ Na, Nc }

2. (Di){ Na, Nc }(DE)

3. (之){ VH }

4. (P, Da, Dk, D, Neqa, DM){ VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, SHI }

5. { Da, Dk, D, Neqa }(VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK,

VL, V_2, SHI)

6. (Dfa){ VH, VI, VJ, VK, VL }

7. { Dfa }(VH, VI, VJ, VK, VL, VE, D)

8. (VH, VI, VJ, VK, VL){ Dfb }

9. { VA, VCL }(在 , 至 , 自 , 離 , 經)

10. (VA , VCL){ 在 , 至 , 自 , 離 , 經 }

11. { Na, Nc, Ncd, Nd, Neu, Nes, Nep, Neqa, Neqb, Nf, Ng, Nh, VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, SHI, DM }(Ng, Ncd)

12. (Na, Nc, Ncd, Nd, Neu, Nes, Nep, Neqa, Neqb, Nf, Ng, Nh, VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, DM) { Ng , Ncd }

Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 45-60
©Computational Linguistics Society of R.O.C.

Meaning Representation and Meaning Instantiation for Chinese Nominals[1]

Kathleen Ahrens[*], Li-li Chang[+], Ke-jiann Chen[+], Chu-Ren Huang[+]

Abstract

The goal of this paper is to explicate the nature of Chinese nominal semantics, and to create a paradigm for nominal semantics in general that will be useful for natural language processing purposes. We first point out that a lexical item may have two meanings simultaneously, and that current models of lexical semantic representation cannot handle this phenomenon. We then propose a meaning representation that deals with this problem, and also discuss how the meanings involved are instantiated. In particular we posit that in addition to the traditional notion of sense differentiation, each sense may have different meaning facets. These meaning facets are linked to their sense or to other meaning facets through one of two ways: meronymic or metonymic extension.

1. Introduction

Lexical ambiguity resolution is a central concern of natural language processing [Small et al., 1988]. The traditional way of looking at the problem is to list the various meanings that a word has, and write a rule-based program to pick the appropriate meaning for the context. Both Categorical Grammar and Montague Semantics, for example, assume that meanings are discrete and that there is a one-to-one correspondence between a lexical item and its meaning translation. The discrete meaning hypothesis provides the conceptual basis for most of the previous literature on ambiguity resolution and semantic resolution. In short, ambiguity resolution is viewed as trying to choose from several discrete meanings that share the same linguistic form (i.e. lexical form). While this approach can provide an algorithm to identify an appropriate meaning in a given context, it cannot account for novel uses of lexical items.

* National Taiwan University, Taipei, Taiwan, R.O.C. E-mail: ahrens@ms.cc.ntu.edu.tw
+ Academia Sinica, Nankang, Taipei, Taiwan, R.O.C.

[1] This paper is jointly authored. The names of the authors are listed in alphabetical order.

More recent work addresses this problem. Pustejovsky's [1995] Generative Lexicon provides a framework (i.e. qualia structure) for possible meanings, and discusses under what conditions which meaning will be chosen (i.e semantic coercion). His account is especially useful in dealing with the creative use of words in novel contexts, an area that had been previously ignored due to the assumption that either a) the novel usage could be listed if necessary, and b) often it was deemed not necessary to list these novel meanings because they occurred so rarely.

However, one issue that Pustejovsky and others have yet to account for is the fact that lexical meaning can be **actively complex**. All models of lexical ambiguity resolution assume that only one solution exists in a given context. In fact, what we will show is that more than one meaning can co-exist in the same context. A lexical item is actively complex if it allows simultaneous multiple interpretations. We will propose a meaning representation for lexical items that captures this complexity.

In addition, although Pustejovsky provides the framework to exclude the possible meanings, he cannot predict the relationship among the meanings, nor allow for cases where different meanings seem to exist simultaneously. Within the general theory of the Generative Lexicon, Copestake and Briscoe [1995] deal with meaning extension by either underspecification or lexical rules, which also implies that only one meaning can be expressed at any given time.

In our account, we will demonstrate that meaning can be predicted from its context by the interaction of a) the semantic class of the item, and b) its possible meaning extensions. Our account has the advantage of being able to account for a wider range of linguistic data, including puns and polysemous uses, in addition to novel extensions. Our account also has the advantage of being both computationally parsimonious, as well as conceptually intuitive.

Our paper is divided as follows: in section 2, we will first present background information and definitions concerning the different kinds of ways that meanings can vary. In section 3, we will present our arguments for the active complexity of lexical meaning, present a representation that can handle active complexity, and also give reasons for the conceptual intuitiveness of the model. In section 4, we will discuss the meaning extensions that have been found to date. Section 5 discusses the hierarchical information that is passed from a semantic class to an individual item of that class. Section 6 summarizes our findings and suggests future areas of research.

2. Background

In this paper we devise a meaning representation for nominals (and Chinese nominals in particular) such that all meaning aspects of a noun are dealt with parsimoniously. Nouns, at first glance, do not seem to warrant representational complexity. When one is asked to think of a noun, one commonly thinks of a concrete object, such as 'paper'. When asked to define it, one could reply that it is a thin, white, rectangular object (appearance) made from the pulp of trees (origin) that people nowadays use to write and print on (function). But 'paper', even if we do not talk about its additional meanings in compound items such as 'wrapping paper', 'tissue paper' etc., has a variety of meanings including: a piece of paper, a newspaper, the office where a newspaper is written, and an academic paper. This phenomenon is not language specific. For example, in Mandarin Chinese, the word 雜誌 'magazine' can refer to the physical object (1a), or the information contained within (1b), or the publishing house (1c).

(1a) 他 手 上 拿 了 本 雜誌。
ta shou shang na le ben zazhi
he hand on hold asp. CL magazine
'He is holding one magazine in his hand.'

(1b) 我們 從 雜誌 中 得到 許多 寶貴的 資料。
women cong zazhi zhong dedao xuduo baoguide ziliao
we from magazine within obtain many precious data
'We have obtained a lot of precious data from magazines.'

(1c) 美國 各 大 雜誌 無不 挖空 心思 爭取 採訪 機會。
meiguo ge da zazhi wubu wakong xinsi zhengqu caifang jihui
America every big magazine do dig-empty mind fight fro interview chance
'Major American magazines fight for interview opportunities.'

Nor is this phenomenon limited to words relating to items that may contain information such as papers and magazines. The word 天 'tian' in Chinese can refer to the sky (2a), God (2b), weather (2c), time (2d), day(s) (2e), or nature (2f). The word 刀 'dao' can refer either to the whole knife (3a), or only to the blade of a knife (3b). The word 梅花 'meihua' can refer either to the plum-flower blossom (4a), or the plum-flower plant (4b). The word 白菜 'baicai' can refer to either the round raw vegetable (5a), or the soft cooked mass (5b).

(2a)	抬頭	望	著	湛藍的	天。
	taitou	wang	zhe	zhanlande	tian
	raise head	watch	asp.	blue	sky

'Raise one's head and look at the blue sky.' ('Tian' refers to sky.)

(2b)	中國人	說	福	自	天	來。
	zhongguoren	shuo	fu	zi	tian	lai
	Chinese	say	happiness	from	sky	come

'Chinese say, happiness comes from heaven.' ('Tian' refers to God/heaven.)

(2c)	天	冷	時	別	忘	了	加	件	衣服。
	tian	leng	shi	bie	wang	le	jia	jian	yifu
	sky	cold	time	not	forget	asp.	add	CL	clothes

'Don't forget to put on more clothes when the weather is cold.'
('Tian' refers to weather.)

(2d)	天	不	早	了。
	tian	bu	zau	le
	sky	not	early	particle

'It is not early.' ('Tian' refers to time.)

(2e)	他	在	這裡	待	了	一	整	天。
	ta	zai	zheli	dai	le	yi	zheng	tian
	he	in	here	stay	asp.	one	whole	sky/day

'He has stayed here for one whole day.' ('Tian' refers to day(s).)

(2f)	人類	是	大部分	動物	的	天敵。
	renlei	shi	dabufen	dongwu	de	tiandi
	human being	is	most	animal	's	natural enemy

'Human beings are the natural enemy of almost all animals.' ('Tian' refers to nature.)

(3a)	我	向	他	借	了	一	把	刀。
	wou	xiang	ta	jie	le	yi	ba	dao
	I	from	him	borrow	asp.	one	CL	knife

'I borrowed a knife from him.' ('Dao' refers to the whole cutting instrument.)

(3b)	這	把	刀	很	利。
	zhe	ba	dao	hen	li
	this	CL	knife	very	sharp

'The knife is very sharp.' ('Dao' refers to the cutting edge.)

(4a)	一	朵	梅花
	yi	duo	meihua
	one	CL	plum-flower

'a plum-flower blossom' ('Meihua' refers to the blossom.)

(4b)	一	棵	梅花
	yi	ke	meihua
	one	CL	plum-flower

'a plum-flower plant' ('Meihua' refers to the whole plant.)

(5a)	一	棵	白菜
	yi	ke	baicai
	one	CL	Chinese cabbage

'a Chinese cabbage' ('Baicai' refers to the vegetable plant.)

(5b)	一	盤	白菜
	yi	pan	baicai
	one	CL	Chinese cabbage

'a dish of Chinese cabbage' ('Baicai' refers to the cooked dish.)

The examples we have given above are all examples of polysemy, which is when a word has several, related meanings. But meanings can also be unrelated, as in the case of the two meanings for 'bank' (i.e. 'financial institution' and 'land on the side of a river'). A noun that has two unrelated meanings is referred to as homonymous. Meanings for a word can also be vague or underspecified. An example of this in English is 'aunt' which can refer to someone's parent's sister, where the gender as to the parent is unspecified. (The parent's gender in other languages, such as Mandarin, is important and specified.) The difference as to whether a word is ambiguous or polysemous depends on the perceived relationship (or lack thereof) between the meanings. The distinction between vagueness and polysemy 'involves the question whether a particular piece of semantic information is part of the underlying semantic structure of the item, or is the result of a contextual (and hence pragmatic) specification' [Geerarts 1993:228].

This definition, however, cannot be applied as straightforwardly as it appears. Consider example (1) above. It could be the case that there is no underlying semantic structure for the three meanings (that is, they are vague), and that context alone 'brings out' these meanings. But 1) intuitively these meanings seem to have an underlying structure, and 2) nouns of a similar semantic class (i.e. magazines and newspapers) have similar meanings, which indicates that an underlying structure exists. If it is the case that the pieces of semantic information are part of the underlying structure of the item, then we must deal with the paradoxical situation (given the definition above) that these different meanings are brought out in different contexts.

Tuggy [1993] points out that ambiguity, polysemy and vagueness are better dealt with on a continuum, rather than as sets with discrete boundaries. The prototypical case of ambiguity is where well-entrenched and salient semantic structures are associated with the same phonological representation, and there is no clear subsuming semantic schema. The prototypical case of vagueness is where the meanings are not well-entrenched, and there is a clear subsuming semantic schema (as in the case of parent's sister for 'aunt'). Polysemy is viewed as being in between these two extremes, with there are well-entrenched and salient semantic structures associated with the same phonological representation, but there is also a subsuming schema.

3. Meaning Representation

3.1 Active Complexity of Lexical Items

The above discussion has assumed that one meaning is chosen in a given context. But that is not necessarily the case. There are two types of active complexity in natural language. The first is 'triggered complexity' and involves puns. For example, in (6) either liquor and shipyard is possible as the meaning of port, but it is also possible for both meanings to exist at the same time.

(6) After the accident, the captain went straight for the port.

Example (6) can mean that a) the captain went straight for shore (but humorously implies that the captain was so shook up as to need a drink), or b) that he went straight for his bottle of liquor and also towards the shore (although this is much less likely since this interpretation is not seen as humorous).

The phenomenon in example (6) is a pun. Puns are a humorous play on ambiguous words. Because puns are used for special linguistic purposes (such as humor), and because it is the effect of co-existing meanings that creates the humor, this phenomenon has not previously been considered to be relevant to lexical semantic analysis and lexical representation. The complexity is triggered since it must be initiated by the speaker.

Second, in Chinese, nouns can be actively complex, even when there is no pun or vagueness intended. This is 'latent complexity.' In (7), for example 'book' must be understood as both a physical object, and as information.

(7) 張三 在 翻閱 那 一 本 書。
Zhangsan zai fanyue na yi ben shu.
Zhangsan PROG turn page/read that one CL book
'Zhangsan is turning the pages of the book and reading it.'

In fact, such latent complexity also exists in English nominal semantics. It is well-known that words referring to building apertures, such as door or window are often lexically ambiguous with the structure built to block that aperture. Thus, door in (8) could only refer to the structure, while door in (9) can only refer to the aperture. However, both the aperture and structure's meanings exist simultaneously for both the English and Chinese sentence in (10).

(8) The door is heavy.

(9) John walked in the door.

(10) 門 很 寬
men hen kuan
door very wide
'The door is very wide',

We think this kind of data presents the strongest argument against representing nominal semantics as discrete meaning translations, and for representing nominal semantics as structured meanings connected by conceptual links, such as the qualia structure in Pustejovsky's Generative Semantics. However, since we have shown that different but related meanings can coexist in the same context, Pustejovsky's formulation where related meanings are represented as different attribute value pairs in a feature matrix is inadequate since only one attribute value pair can be picked in each context. We posit that these related meanings are like the facets of a three-dimensional object, such as a diamond, where the meaning instantiation could be a straightforward single facet or multiple connected facets, depending on the context.

3.2 Meaning Representation

The meaning representation that we select is quite straightforward, but differs from other representations in several crucial respects. First, words are listed (following Chinese lexicographic tradition) in terms of their orthographic representation (i.e. the stroke order of the Chinese characters.) Then the senses for each word are listed. The phonological representations are associated with each sense listing, and may or may not be the same. Second, the sense differentiation includes senses that are related (polysemous senses) as well as unrelated (homonymous senses). There is no attempt in this representation to distinguish clearly between those meanings that are polysemous or homonymous. This is because speakers tend to draw their own conclusions about the relationships between senses (e.g. many speakers see a relationship between 'ear of corn' and 'ear that you hear with', although there is no historical or semantic relationship whatsoever [Lyons 1977]).[2] Third, and most importantly, our lexical representation has **meaning facets** located within each sense. Meaning facets reflect an aspect of a sense. For example, in (11)-(14) we show examples of words with one sense, of which there are two to three different meaning facets.

(11) 雜誌 **--Sense$_1$: ZAZHI *magazine*** -- meaning facet$_1$: *physical object*
-- meaning facet$_2$: *information contained within*
-- meaning facet$_3$: *institution that publishes magazine*

(12) 刀 --**Sense$_{1:}$ DAO *knife*** -- meaning facet$_1$: *physical object*
-- meaning facet$_2$: *the blade of it*

[2] However, if a study was run on native speakers to find out their understanding of the relative closeness of relationship among meanings, this information could be incorporated into our representation by simply indicating which senses should be grouped together.

(13) 梅花 **--Sense$_1$: MEIHUA *plum flower***
-- meaning facet$_1$: *physical object: the blossom*
-- meaning facet$_2$: *the whole plant contains the blossom*

(14) 白菜 **--Sense$_1$: BAICAI *Chinese cabbage***
-- meaning facet$_1$: *physical object: the vegetable*
-- meaning facet$_2$: *the cooked form of it*

In (15) we give an example of a word with four different senses, of which one has three different meaning facets.

(15) 天 **--Sense$_{1:}$ TIAN *sky*** -- meaning facet$_1$: *sky as a physical object (that can be viewed)*
-- meaning facet$_2$: *God/heaven*
-- meaning facet$_3$: *weather*

--Sense$_2$: TIAN *time*
--Sense$_3$: TIAN *day*
--Sense$_4$:TIAN *nature*

How do we decide whether a certain meaning is a sense or a meaning facet? A meaning facet is an extension from a particular sense. It has the following three properties: 1) it can appear in the same context as other meaning facets, 2) it is an extension from a core sense or from another meaning facet (unless it is the core sense), 3) nouns of the same semantic classes will have similar sense extensions to related meaning facets. Individual senses, on the other hand, 1) cannot appear in the same context (unless the complexity is triggered), 2) have no core sense from which it is extended, or it is very hard to concisely define what the core sense would be, and 3) no logical/conceptual links can be established between two senses, non can the link between two senses be inherited by class of nouns.

For example, in (16) below, we can see that the meaning of sky (as a physical object) and God can appear in the same context, as can sky (as a physical object) and weather (17), sky (as a physical object), God, and weather (18). Thus, they are all different meaning facets of the first sense in (15).

(16) 有 人 開始 不 敬 天 也 不 拜 天 了。
you ren kaishi bu jing tian ye bu bai tian le
there are person begin not respect sky and not worship sky particle
'There are people who ceased to respect heaven or to worship heaven.'
('Tian' refers to both sky and God/heaven.)

(17) 天 放晴 了。
tian fangqing le
sky become sunny particle
'It became sunny.' ('Tian' refers to both sky and weather.)

(18) 農民 長久 靠 天 依 地 的 生活。
nongmin changjiou kau tian yi di de shenghuo
farmer long depend sky depend ground DE live
'Farmers have long lived a life that depends on heaven and earth.'
('Tian' refers to sky, God, and weather.)

The above examples also demonstrate that only one sense can occur in any given context. The sense of 'time' or 'day' or 'nature' is not available in any one of the above contexts.[3] Only meaning facets of a particular sense can be available in the same context. Context, in effect, selects which sense is made available. Context may also select a particular meaning facet, as in (2a)- (2c), but it does not necessarily have to, because context may activate several meaning facets at once, as in (16) - (18).

What aspects of context help to pick a sense or a meaning facet? Verbs and prepositions are usually instrumental in determining which meaning can occur in which context. For example, in the above instance, the meaning of 'God' can only occur with volitional verbs and cannot occur with verbs having to do with pure locative. The type of contextual information that picks out one sense or one meaning facet is an important area of future research.

[3] 'Time' might be viewed as a meaning facet of the sense 'sky', as shown by the identical strings in (i) and (ii).

(i) [$_s$ 天 [$_{VP}$ 黑 了]]。
tian hei le
sky dark particle
'The sky turned dark.'

(ii) [$_S$[$_{VP}$[$_V$ 天 黑] 了]]]。
tianhei le
sunset particle
'The sun has set (i.e. it is late).'

However, the interpretation in (i) is a subject-predicate sentence, while the interpretation in (ii) involves a disyllabic lexical item. Thus, these two sentences are structurally different and no latent complexity is involved.

3.3 Conceptual Adventages

Viewed from this perspective, context always plays a role in determining which meaning is chosen, whether the word is ambiguous, polysemous, or vague. Tuggy's meaning models were two dimensional. But we suggest that a 3-dimension model allows for a greater understanding of the relationship between meaning and context. Imagine a multi-faceted object, such as a cube. Imagine that there is a core in the center of the cube, and that there are lines that radiate out to each of the six surfaces (i.e. this would be the case for a word that had six senses). The core represents the orthographic representation of the word, and each surface represents a different sense of the word and its associated phonological representation (i.e. the information that is bolded in our lexical representation above). Furthermore, from each surface of the cube, there may also be (dotted) lines that radiate out to additional surfaces, which are the facets of that particular sense (i.e. the non-bolded information in our lexical representation above). Thus, when context turns the cube so that one particular sense surface is shown to a light source (i.e. the hearer) then light is reflected from only that surface, and only that sense is computed. In the case, however, where context turns the cube so that a sense surface that has meaning facets extending from it is shown to a light source, the light can reflect off of any one, or any combination of the meaning facets, just as light can reflect from the different facets of a diamond. Our representation, then, is not only computationally adequate, it is also conceptually intuitive.

In what follows we present the types of links that can occur in noun meaning representations, and we also present the underlying schema for the information contained in each meaning facet.

4. Meaning Links

In our model the meaning representation is structured, and the structure is built upon meaning links. One implication of this model is that semantic classes in a semantic hierarchy will inherit both traditional semantic features as well meaning link structures. Lexical semantic issues will therefore be defined in terms of 1) lexical senses, 2) the possible meaning links of their sense classes and 3) constraints on meaning extensions through these links.

The relationship between a sense and its meaning facets is an area that deserves in depth research and analysis. What follows is a preliminary report of our findings to date. We have found that there are two main ways that meaning facets can extend either from a sense or from another meaning facet: meronymic and metonymic extensions.

4.1 Meronymic extensions

Meronymic extensions involve both the whole standing for part, and part standing for whole. We observe that meronymic extensions are driven by cognitive and conceptual saliency. For example, in (3b) knife actually refers to the blade of the knife. This meronymic extension is motivated by the fact that 'blade' is the locus of cutting, and the most salient function of knife. We also observe that such cognitively driven extensions are not sensitive to blocking effects. For instance, the instance of the specific term 刀刃 'blade' does not block us from saying 'the knife is sharp' as in (3b). Our speculation here is that only conventionalized usages are subject to blocking effects since blocking is the result of (competing) conventions.

In the case of part standing for whole, cognitive saliency is again the prime motivator of the extension. For example, in the case of (19), plum-flower stands for the whole plum tree. The plum flower with its color and scent and endurance in cold weather is the most cognitively salient aspect of the plum tree (for Chinese).

(19) 院子 裡 有 許多 梅花

yuanzi li you xuduo meihua
garden inside exist many plum-flower
'There are many plum-flowers in the garden.'

4.2 Metonymic Extensions

Metonymic extensions are different from meronymic extensions in that the extended meaning is related to the origin of the basic sense, but is not inherent to the basic sense (cf. the part-whole relation above). Metonymic extensions are typically driven by certain eventive relationships such as the ones encoded in Pustejovsky's qualia structure. Unlike meronymic extensions, metonymic extensions are often sensitive to blocking effects. For instance, the grinding extension allows the individual terms to refer to a mass produced from that individual. For example, in (5b) the basic meaning ' 白菜 baicai' refers to the cabbage plant, but after the grinding extension it refers to a mass noun. But in the case of rice ' 米 mi', the grinding extension does not work, because there is a term ' 飯 fan' (cooked rice) already.

4.3 Partial list of Meaning Links

We give here a partial list of the meaning links found to date. We also provide the list of semantic classes that we have found to inherit these links.

I. Meronymic Extensions

1. Whole for part
    - a. whole → functional part {semantic class: artifacts, buildings}
    - b. whole → sentiently salient part {semantic class: body parts}
2. Part for whole
    - a. conceptually salient part → whole {semantic class: fruit, flower}

II. Metonymic Extensions

1. agentivization
    - a. information media → information creator {semantic class: publications}
2. product instantiation
    - a. institution → product {semantic class: manufacturer, trademarks}
3. grinding
    - a. individual → mass {semantic class: vegetables, fruits}
4. portioning
    - a. information media → information {semantic class: publications}
    - b. container → containee
    - c. body part → function
5. space mark-up
    - a. landmark → space in vicinity {semantic class: locations, landmarks}
    - b. structure → aperture {semantic class: doors, windows}
    - c. institution → locus {semantic class: institutions}
6. time mark-up
    - a. event → temporal period
    - b. object → process
    - c. locus → duration

A summary of the links used in the lexical representation of the words we define in this paper is given below (cf. ex. 11): First, the meaning links between the different facets of 'zhazhi' (magazine) are as follows: the first meaning link refers to the concept of magazine as a physical object, the second meaning link is a metonymic extension that relates information media to information, and the third meaning link is a metonymic extension that relates information media to information creator. The link between the two facets of 'dao' (knife) (cf. ex. 12) are that the first link refers to the concept of knife as a physical object (in its entirety), and the second link is a meronymic extension (whole for part) to the meaning facet of 'blade.' The link between the two facets of 'meihua' (plum-flower) (cf. ex. 13) are that the first link refers to the conceptually salient notion of plum-flower, and the second link is a meronymic extension (part for whole) to the meaning facet of plum tree. The link between the two facets of 'baicai' (cabbage) (cf. ex. 14) is that the first link refers to the individual head of cabbage, and the second link is a metonymic extension (grinding) to the meaning facet of a 'dish of cabbage'. The links among the facets of 'tian' (sky) (cf. ex. 15) are that the first link refers to sky as a physical object, the second link is a metonymic extension of space mark-up, and the third link is a meronymic extension of whole extending to the sententially salient part.

We have found that these two types of links (i.e. meronymic and metonymic extensions) are the most productive among meaning extensions. This might be because these types of extensions refer only to the knowledge concerning the lexical item itself. Metaphorical extensions, on the other hand, map a domain of knowledge that does not have anything to do with the lexical item onto the domain of knowledge surrounding the lexical item. Thus, metaphorical extensions are clearly conceptually more complex than metonymic and meronymic extensions, and will be the focus of future research.

5. Meaning Inheritance

Another important issue in lexical semantics is the semantic class. Traditionally, the taxonomic hierarchies are discussed in terms of ISA relationship and inherited features, such as humanness and animacy [Chen and Cha 1988, Sowa 1993]. However, this simplistic traditional model (such as Schank's well-known semantic network) have difficulties when certain nodes do not necessarily inherit all the features from the higher nodes. For example, an ostrich is a bird, but it cannot inherit the feature of [+flight] because it does not fly. Default override of inheritance is computationally plausible though costly.

The other problem with traditional semantic hierarchies has to do with multiple inheritance. For instance, it is intuitive to classifiy ' 籃球 lan-qiu' (basketball) as a physical object. However, it is also clearly an abstract event (i.e. the basketball game). Hence there is cross-taxonomic paradox, which is usually accounted for with the computationally costly mechanism of multiple inheritance [Briscoe et al., 1993].

In our model, both kinds of inheritance problems disappear since what a semantic class shares is a partial structure of semantic links. That is, we will annotate meaning links to a semantic class, and these links will be inherited by all the members of the class.[4] In the case of ' 球 qiu' (ball), it inherits the metonymic link of a round physical object and extends to the game played with the object. This explanation is more parsimonious since it reduces the costly computation of multiple inheritance and makes most

[4] Of course, the lexicon would have to specify any blocking effects where the linking does not apply.

cases of the local overriding of inheritance unnecessary. It is also conceptually powerful in allowing richer semantic representation. For instance, the semantic class of flowers will inherit the meronymic extension of part for whole.

6. Conclusion: Implementation and Implications

Traditional methods of dealing with ambiguity and vagueness in natural language processing have been complicated by the on-line compilations that are usually necessary to deal with the 'additional' meanings created by the context. But our account postulates multiple senses and structured ways of linking additional meaning facets to the senses so that the information is all listed in the representation, and therefore easier to access. Our proposal is to have not only the different senses of a word listed, but also its different meaning facets. We claim that there are conceptual or logical relationships between the facets and their senses, as discussed in section 4.

The organization that we have proposed here is a shallow structure, with only two levels: the sense level and the meaning facet level. Both levels can be annotated with meaning links. Conceptually it is as explanatory as a theory where all the meaning links are structurally represented. This is because all represented meaning links can be traced, and a (semantic-class-based) meaning derivation tree can be established off-line. Moreover, not having an overt tree of meaning extensions allows us to avoid multiple-inheritance and blocking problems. A shallow structure also allows efficient access, reflecting the psychological reality that the depth of meaning derviation is not relevant in lexical access.

In this paper we propose a meaning representation for Chinese nominal semantics, as well as a paradigm for nominal semantics in general that will be useful for natural language processing purposes. We point out that a lexical item may have two meanings simultaneously, which current models of lexical semantic representation cannot handle. We call this phenomena 'active complexity.' There are two types of active complexity: 'triggered complexity' where the noun is purposely selected to be simultaneously ambiguous, and 'latent complexity' where the noun selected has two or more meanings coexist, but the effect is not humorous and was not selected for such an effect. We propose a meaning representation to account for this phenomena, and also discuss how the meanings involved are instantiated. We postulate that in addition to the traditional notion of sense differentiation, each sense may have different meaning facets. These meaning facets link to their sense or to other meaning facets through one of two ways: meronymic or metonymic extension. We also point out that instead of a traditional taxonomic relationship, what is being inherited in addition to semantic features is meaning extensions/relations, such that words of the same semantic class have the same meaning extensions. Our representation, therefore, allows for predictions of meaning extensions from a semantic class.

The representation proposed here is the result of extensive corpus-based studies of the 200 most productive nominal endings in Mandarin [CKIP 1995]. These productive nominal endings in turn each derive scores of highly frequent nouns. Hence we have accounted for a substantial portion of Chinese noun usages. We have also provided detailed semantic representation of the nominal heads based on our proposed representation. This is a significant first step towards the comprehensive formal representation of Mandarin nominal semantics and is also the first step towards fully automated Mandarin Language Understanding.

References

Briscoe, E., J. Copestake, and V. de Paiva, *Inheritance, Defaults and the Lexicon*. Cambridge University Press, 1993.

Chen, K.-J., and C.-S. Cha, "The Design of a Conceptual Structure and Its Relation to the Parsing of Chinese Sentences," *Proceedings of 1988 International Conference on Computer Processing of Chinese and Oriental Languages* (*ICCPCOL*), 1988, pp. 428-431.

CKIP, "Contents and Explanations of Sinica Corpus," *CKIP Technical Report. 95-02*. Nankang: Academia Sinica, 1995.

Copestake, A., and T. Briscoe, "Semi-productive Polysemy and Sense Extension," *Journal of Semantics*, 12, 1995, pp.15-67.

Geerarts, D., "Vagueness's puzzles, polysemy's vagaries," *Cognitive Linguistics*, 4.3, 1993, pp. 223-272.

Lyons, J., *Semantics*, Cambridge University Press, 1977.

Pustejovsky, J., *The Generative Lexicon*, MIT Press, 1995.

Small, S., G. Cottrell, and M. Tanenhaus, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence*, Morgan Kaufmann Publishers, 1988.

Sowa, J., "Lexical Structure and Conceptual Structure," in *Semantics and the Lexicon*, Pustejovsky (Ed.), Kluwer Academic Publishers, 1993, pp. 223-262.

Tuggy, D., "Ambiguity, Polysemy and Vagueness," *Cognitive Linguistics*. 4.3, 1993, pp. 273-290.

Computatioal Linguistics and Chinese Language Processing
vol 3, no. 1 , February 98, pp. 61-74
©Computational Linguisitcs Society of R. O. C.

Towards a Representation of Verbal Semantics -- An Approach Based on Near-Synonyms

Mei-Chih Tsai*, Chu-Ren Huang*, Keh-Jiann Chen*, Kathleen Ahrens+

Abstract

In this paper we propose using the distributional differences in the syntactic patterns of near-synonyms to deduce the relevant components of verb meaning. Our method involves determining the distributional differences in syntactic patterns, deducing the semantic features from the syntactic phenomena, and testing the semantic features in new syntactic frames. We determine the distributional differences in syntactic patterns through the following five steps: First, we search for all instances of the verb in the corpus. Second, we classify each of these instances into its type of syntactic function. Third, we classify each of these instances into its argument structure type. Fourth, we determine the aspectual type that is associated with each verb. Lastly, we determine each verb's sentential type. Once the distributional differences have been determined, then the relevant semantic features are postulated. Our goal is to tease out the lexical semantic features as the explanation, and as the motivation of the syntactic contrasts.

1. Introduction

Radical Lexicalism maintains that all grammatical behaviors are manifestations of lexical features [Karttunen 1986]. Since most lexical attributes are semantic and/or conceptual in nature, taking this lexicon-driven approach to language means that many syntactic properties can be predicted from lexical semantic attributes [Jackendoff 1976, Levin 1985, Dowty 1991, Pustejovsky 1993]. In terms of Natural Language Processing (NLP), surface syntactic structures can be systematically predicted from their lexical semantic representation. From this perspective, the automatic acquisition of lexical knowledge for NLP may be possible, since the relation between syntactic patterns and lexical semantics is predictable to some extent. Dorr & Jones [1996], for example, demonstrate that semantic information can be derived from syntactic cues when the syntactic cues are first divided into distinct groupings that correlate with different word senses.

* Academia Sinica, Nankang, Taipei 115, Taiwan, R.O.C. E-mail: tsmei@mail.nsysu.edu.tw
+ National Taiwan University, Taipei 106, Taiwan, R.O.C.

However, as Levin [1993] points out, there are still many questions to be explored:

> If the hypothesis that syntactic properties are semantically determined is taken seriously, then the task is to determine, first, to what extent the meaning of a verb determines its syntactic behavior, and second, to the extent that syntactic behavior is predictable, what components of verb meaning figure in the relevant generalizations. The identification of the relevant components of meaning is essential if this approach is to be successful.[Levin 1993:14]

Our paper will focus on the last point above. We propose using the distributional differences in the syntactic patterns of near-synonyms[1] to deduce the relevant components of verbal semantics. In particular, we want to identify the semantic features that differentiate verbal syntactic behaviors. Our strong hypothesis is that all lexical semantic features can be identified this way. In contrast, salient semantic features deduced from a shared verb class may or may not be predictive of verbal features because they may simply be descriptions of the meaning. Our method is as follows:

1) Determine distributional differences in syntactic patterns

2) Deduce the semantic features from the syntactic phenomena

3) Test the semantic features in new syntactic frames

How will we determine the distributional differences in syntactic patterns? Our corpus-based approach calls for us to search, sort, and classify all relevant data according to the four following criteria: First, we will classify each of these instances according to the syntactic functions of the verbs themselves (i.e. predicate, complement, adverbial, determinal, nominal). Second, we can classify the corpus data in terms of argument type that the verbs take (i.e. NP subject, VP subject, sentential subject, NP object, NP double-object, sentential object). Third, we determine the aspectual types each verb is associated with (i.e. aspectual markers, aspectual adverbs, resultative complements). Lastly, we examine the sentential modes that each verb occurs in (i.e. passive, imperative, evaluative, declarative, interrogative).

[1] According to Lyons (1995: 60), synonyms are expressions with the same meaning, whereas near-synonyms are expressions that are more or less similar, but not identical, in meaning. In this respect, many of the expressions listed as synonymous in dictionaries, actually, are near-synonyms.

This process is time-consuming. However, because we are dealing with near-synonyms, we expect there to be many shared syntactic behaviors that can be ignored for the purpose of this study. This will facilitate the identification of (sometimes unexpected) grammatical contrasts that instantiates deeper lexical semantic contrasts of the near-synonym pairs. The crucial difference will be found in the small number of instances where they are in complementary distribution in terms of one of the above four types of syntactic information.[2] In what follows we will present our 3 - step methodology (i.e. determine syntactic difference, deduce semantic feature, test for reliability of semantic feature) for each of the 4 different types of syntactic information (i.e. syntactic functions (Section 2), argument structure (Section 3), aspectual type (Section 4), sentential type (Section 5). In the concluding section (Section 6), we discuss the advantages of this method as compared to an account that is based on differentiating semantic classes of verbs [Levin 1993].

2. Syntactic functions

In this section, we look at what type of syntactic functions a verb can occur with, including predicate, adverbial, complement, nominalization, etc.

2.1 Distributional differences

The distributional contrasts in terms of the syntactic functions between the two state verbs LEI 'be tired' and PIJUAN 'be tired' are that LEI functions as a (resultative) complement in 6% of the cases, but never occurs in a nominal phrase, while PIJUAN serves as a noun in 9% of the instances, but never occurs in a (resultative) complement position. The data from the Academia Sinica Balanced Corpus[3] (abb. Sinica Corpus) is given in Table 1 and the relevant examples are given in (1) and (2). (The numbers next to the verbs in the table indicate the number of instances of occurrence in the entire Sinica Corpus.)

Functions	Complement	Nominalization
LEI 174	11 (6%)	--
PIJUAN 33	--	3 (9%)

Table 1. *Table 1. Differences in syntactic functions: LEI vs. PIJUAN*

[2] Due to space limit, only three pairs of verbs are illustrated in this paper: LEI-PIJUAN, 'be tired', QUAN-SHUIFU, 'persuade', GAOXING-KUAILE, 'be happy'. The last pair GAOXING and KUAILE are not included in Teng's synonyms dictionary because their corresponding terms in English are different.

[3] Academia Sinica Balanced Corpus is the largest balanced corpus of both written and spoken contemporary Mandarin, developed by CKIP group in Academia Sinica, Taiwan, containing 3.5 million words.

(1) Resultative complement

(1a) ta zou de hen lei[4]
he walk DE very be-tired
'He walked so much that he was tired.'

(1b) # ta zou de hen pijuan
he walk DE very be-tired

(2) Nominalized object

(2a) shuimian shi zhi pijuan zuihaode fangfa
sleep be treat be-tired best method
'Sleeping is the best method to treat the tiredness.'

(2b) # shuimian shi zhi lei zuihaode fangfa
sleep be treat be-tired best method

2.2 Semantic feature

One semantic feature that would distinguish the meaning of these two verbs is [+/-effect]. In other words, though both are states that predicate of people, LEI has the additional meaning that is an effect state of an (unspecified) event, while PIJUAN does not specify this. It is obvious that an effect state occurs as a resultative complement, and represents the effect of another predicate. On the other hand, there seems to be a tendency against nominalized complex verbs in Chinese (e.g. all verb-resultative compounds cannot be nominalized). Thus, an effect state has the semantic implicature of a complex event and cannot be nominalized.

2.3 Prediction/Verification

After looking at near-synonyms to determine the semantic feature that differentiates them, we need to test our hypothesis. The following two examples demonstrate that it is much easier for LEI than for PIJUAN to occur with the perfective aspect marker (ASP) *-le*. The statistics shown in Table 2 indicate the relatively high percentage of LEI co-occurring with *-le* when compared with the zero utterance of PIJUAN.

(3) Perfective aspect marker

(3a) tamen lei le jiu lai ci he pijiu
they be-tired ASP then come here drink beer
'When they get tired, they come here to drink some beer.'

[4] The abbreviations used in the glosses are the following: ASP 'aspect maker', BEI 'passive maker,' CL 'classifier,' PAR 'sentential-final particle.' Examples begin with a # are either unnatural or inacceptable.

(3b) # tamen pijuan le jiu lai ci he pijiu
they be-tired ASP then come here drink beer

Collocation	*-le*
LEI 174	38 (22%)
PIJUAN 33	--

Table 2. *Differences in collocations: LEI vs. PIJUAN*

According to Smith [1991], perfective *-le* appears only in dynamic sentences, presenting closed non-stative situations. When stative verbs occur with this morpheme, the sentences have only inchoative reading with focus on the initial point of the state. Thus, the collocation to *-le* reveals that the state expressed by LEI results in a change of state. In other words, LEI is an effective state, i.e. [+ effect]. PIJUAN, on the other hand, is a genuine state, i.e. [- effect].

In addition to the perfective aspect marker, LEI and PIJUAN also differ in the association with durational complements. While LEI takes a durational complement in 2% of the cases, PIJUAN never does.

(4) Durational complement

(4a) tamen lei le yi xiawu
they be-tired ASP one afternoon
'They have tired themselves all afternoon.'

(4b) # tamen pijuan le yi xiawu
they be-tired ASP one afternoon

Collocation	Durational Complement
LEI 174	4 (2%)
PIJUAN 33	--

Table 3. *Differences in collocations: LEI vs. PIJUAN*

As durational complements are used to locate an interval during which the prediction holds true [cf. Paris 1988], it is expected that there be endpoints in the state LEI. Again the feature [+/- effect] distinguishes the two state verbs in question.

3. Argument selection

The distributional differences for argument selection involve determining whether the verb occurs with an NP subject, VP subject, sentential subject, NP object, double NP object, sentential object, etc.

3.1 Distributional differences

In the case of GAOXING and KUAILE 'be happy', GAOXING can take a sentential object in more than 7% of the cases, while KUAILE cannot, as shown in Table 4 and example (5).

Collocation	Sentential Object
GAOXING 280	20 (7.1%)
KUAILE 365	--

Table 4. *Differences in argument selection: GAOXING vs. KUAILE*

(5) Sentential Object

(5a) tamen hen gaoxing Zhangsan mei zou
they very be-happy John not go away
'They were glad that John did not go away.'

(5b) # tamen hen kuaile Zhangsan mei zou
they very be-happy John not go-away

3.2 Semantic feature

The semantic feature that can be deduced from this distributional difference is [+/-effect], where GAOXING is an effect state triggered off by the cause expressed in the sentential object.

3.3 Prediction/Verification

We observe from the data that only GAOXING can be associated with the perfective aspect marker *-le* in 0.7 % of the instances, as demonstrated below.

(6) Perfective aspect marker

(6a) keren gaoxing le jiu gei xiaofei
customer be-happy ASP then give tip
'When customers are pleased, they give tips.'

(6b) # keren kuaile le jiu gei xiaofei
customer be-happy ASP then give tip

Collocation	*-le*
GAOXING 280	2 (0.7%)
KUAILE 365	--

Table 5. *Differences in collocations: GAOXING vs. KUAILE*

The contrast between (6a) and (6b) is correctly predicted, because it is possible for GAOXING to represent a changed state brought out by some cause, but not for KUAILE. It is then justified to say that GAOXING is an effect state, i.e. [+ effect], whereas KUAILE is [- effect].

4. Aspectual types

The distributional difference for aspectual types involve looking at the aspect markers, aspectual adverbs and resultative complements the verbs co-occur with.

4.1 Distributional differences

In the case of QUAN and SHUIFU 'persuade', only QUAN occurs with the durative aspect marker *-zhe*[5] in 1.8% of the cases, SHUIFU never does.

Collocation	*-zhe*
QUAN 112	2 (1.8%)
SHUIFU 50	--

Table 6. *Differences in collocations: QUAN vs. SHUIFU*

(7) Durative aspect marker

(7a) ta yimian zou, yimian quan-zhe Zhangsan
he one-side walk one-side persuade ASP John
'He persuaded John as he walked.'

(7b) # ta yimian zou, yimian shuifu-zhe Zhangsan
he one-side walk one-side persuade ASP John

4.2 Semantic Feature

As the marker *-zhe* indicates that an event is on-going [cf. Li & Thompson 1981], the fact

[5] Some authors consider *-zhe* as imperfective aspect marker [Ma 1985, Smith 1991].

that QUAN can take such a marker and SHUIFU never can suggests that there are aspectual differences between these two verbs. On the one hand, QUAN denotes an extensible, atelic event. On the other hand, SHUIFU denotes a bounded, telic event. The semantic feature that would distinguish the meaning of these two verbs is [+/- telic].

4.3 Prediction/Verification

If our hypothesis is correct, we expect that only QUAN is compatible with adverbs indicating the durative aspect. Consider the following examples.

(8) Durative aspectual adverb

(8a) ta yizhi quan Zhangsan jiehun
he all-the-time persuade John get-married
'All the time he persuaded John to get married.'

(8b) # ta yizhi shuifu Zhangsan jiehun
he all-the-time persuade John get-married

The adverb *yizhi* 'all the time' in the above examples can only occur with QUAN but not with SHUIFU. This means that only the event denoted by QUAN can be in progress. The difference between these two verbs in telicity is then justified.

A second argument in support of the claim that QUAN differs from SHUIFU in verbal aspect is related to the fact that only QUAN admits, in 3.6% of instances, resultative complements which indicate completion or termination [cf. Smith 1991]. Consider the examples in (9).

Collocation	Resultative Complement
QUAN 112	4 (3.6%)
SHUIFU 50	--

Table 7. *Differences in collocations: QUAN vs. SHUIFU*

(9) Resultative complement

(9a) ta quan de Zhangsan xin hen fan
he persuade DE John mood very be-bored
'He kept trying to persuade John until John was bored to death.'

(9b) # ta shuifu de Zhangsan xin hen fan
he persuade DE John mood very be-bored

It is reasonable that telic verbs like SHUIFU exclude the possibility of taking resultative complements, since we cannot terminate an event which is already terminated. But for atelic verbs like QUAN, it is natural that they take resultative complements, indicating that events are accomplished. Thus the feature [+/- telic] can account for the contrastive use of aspectual type between these two items.

5. Sentential types

In this section, we look at what type of sentences a verb can join, including passive sentence, imperative sentence, wish sentence, evaluative sentence, etc.

5.1 Distributional differences

One of the distributional contrasts between QUAN and SHUIFU involves the possibility of forming passive sentence. It seems that SHUIFU occurs more frequently in passive construction (6%) than QUAN does (0.9%). The examples in (10) show that QUAN is not allowed in the passive construction without a resultative complement.

Collocation	Passive Sentences
QUAN 112	1 (0.9%)
SHUIFU 50	3 (6%)

Table 8. *Differences in collocations: QUAN vs. SHUIFU*

(10) Passive sentence

(10a) # Zhangsan bei ta quan le
John BEI he persuade PAR

(10b) Zhangsan bei ta shuifu le
John BEI he persuade PAR
'John was persuaded by him.'

(10c) Zhangsan bei ta quan-zou le
he BEI he persuade go-away PAR
'John was persuaded to leave by him.'

In case of GAOXING and KUAILE 'be happy', the following distributional contrasts in terms of the sentential types are noticed from the Sinica Corpus: GAOXING never constitutes wish sentences but admits evaluational sentences (1.8%), while KUAILE occurs in wish sentences (2.2%) but never appears in evaluational sentences.

Collocation	Wish Sentences	Evaluational Sentences
GAOXING 280	--	5 (1.8%)
KUAILE 365	8 (2.2%)	--

Table 9. *Differences in collocations: GAOXING vs. KUAILE*

(11) Wish sentence

(11a) zhu ni kuaile!
wish you be-happy
'I wish you be happy.'

(11b) # zhu ni gaoxing!
wish you be-happy

(12) Evaluational sentences

(12a) zhei-jian shi zhide gaoxing.
this CL thing be-worth be-happy
'This thing is worth enjoying.

(12b) # zhei-jian shi zhide kuaile
this CL thing be-worth be-happy

5.2 Semantic Feature

The semantic feature that would distinguish the meaning of QUAN and SHUIFU is [+/-effect]. Though both are events, SHUIFU has an additional meaning of effect which corresponds to the affectedness property of passive sentences, while QUAN does not have.

As for GAOXING and KUAILE, the distinctive feature of their meaning is [+/-control]. Though both are states, only the controllable one can GAOXING express the calculated reaction in evaluational sentences and refuses the impredictive nature of wish sentences.

5.3 Prediction/Verification

We have seen in (9) above that it is possible for QUAN but not for SHUIFU to take a resultative complement. This collocational difference constitutes a good argument for the claim that the meaning of QUAN and SHUIFU can be distinguished by the feature of effect. One point needs to be clarified: why QUAN cannot occur in passive sentences alone without a resultative complement behind? Given that resultative complements not only indicate the accomplishment of the main event, but also express the affected state of the participant, then, the use of such elements can contribute to QUAN additional properties like completion and affectedness, which are inherent to SHUIFU.

Now let us turn to the semantic feature [+/- control]. To support the claim that GAOXING can be controlled and KUAILE cannot, consider the use of imperative sentence illustrated below.

Collocation	Imperative Sentences
GAOXING 280	3 (1.1%)
KUAILE 365	--

Table 10. *Differences in collocations: GAOXING vs. KUAILE*

(13) Imperative sentence

(13a) bie gaoxing!
don't be-happy
'Don't be happy!'

(13b) # bie kuaile!
don't be-happy

The data show that GAOXING can form imperative sentences in 1.1% of the instances, while KUAILE never can. This means that the hearer can only change the state of GAOXING, but not the state of KUAILE. In other words, only the state of GAOXING is controllable.

6. Conclusion

The notion that the syntactic behavior of verbs is semantically determined has been examined extensively, especially for English verbs (please see Levin 1993 for relevant references). The technique that has been used quite productively is one that determines the distinctive behavior of verb classes. Levin summarizes this method:

> The assumption that the syntactic behavior of verbs is semantically determined gives rise to a powerful technique for investigating verb meaning that can be exploited in the development of a theory of lexical knowledge. If the distinctive behavior of verb classes with respect to diathesis alternations arises from their meaning, any class of verbs

> whose members pattern together with respect to diathesis alternation should be a semantically coherent class: its members should share at least some aspect of meaning. Once such a class is identified, its members can be examined to isolate the meaning components they have in common. Thus diathesis alternations can be used to provide a probe into the elements entering into the lexical representation of word meaning. [Levin 1993:14]

However, this technique is not easily implemented in Mandarin, because extensive study of diathesis alternations has not been done in Mandarin. Perhaps one reason is because Mandarin allows both subject and object omission, which means that it is very difficult to get a handle on what is a relevant 'alternation.' The work that has been done on semantic interpretations of syntactic structures (and the verbs that may occur in these structures) in Mandarin, such as in the case of pre-posed objects (such as BA and BEI), while interesting, is inconclusive because the wide variety of contexts and possible meanings defies a unified explanation. [Cf. Thompson 1973, Mei 1978, Bennett 1981, Ren 1991, Sun 1995, etc]

Moreover, the diathesis alternation technique does not allow for a very fine grained analysis of semantic features, because verbs may belong to more than one (seemingly unrelated) alternation class[6], and because different verb classes may share the same alternation[7]. Thus, it is difficult to extract the common semantic feature that predict the difference between the classes. When we look at near-synonyms, on the other hand, we are able to set up a controlled study of lexical semantic contrasts and their grammatical effects[8]. We hope that this fine-grained approach will aid us in identifying the semantic features or attributes that dictate the syntactic differences of verbs.

[6] For example, according to Levin (1993), 'hit' belongs to verbs of throwing, verbs of contact by impact as well as verbs of existence, whereas 'cut' belongs to seven classes--verbs of cutting, verbs of separating and dissembling, verbs of creation and transformation, verbs of psychological state, verbs of bodily state and damage to the body, verbs of grooming and bodily care and meander verbs.

[7] For example, 'hit' and 'cut' share the conative alternation.

[8] Effectively, more larger scale experiment would be needed to deduce the semantic features for more verbs as well as to determine to what extent can this approach be generalized.

References

Bennett, P., "The Evolution of Passive and Disposal Sentences," *Journal of Chinese Linguistics* 9, 1981, pp. 61-89.

Dorr, B. J., and D. Jones, "Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues," in *Proceedings of 16th International Conference on Computational Linguistics* (*COLING 96*), 1996, pp. 322-327.

Dowty, D. R., "Thematic Proto-Roles and Argument Selection," *Language* 67, 1991, pp. 547-619.

Jackendoff, R. S., "Towards an Explanatory Semantic Representation," *Linguistic Inquiry* 7, 1976, pp. 89-150.

Karttunen, L., *Radical Lexicalism*, CSLI-86-68.

Levin, B. (Ed.), *Lexical Semantics in Review*, *Lexicon Project Working Papers 1*, Center for Cognitive Science, MIT, 1985.

Levin, B., *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, 1993.

Liu, M., "Lexical Meaning and Discourse Patterning - The three Mandarin cases of 'build'," paper presented in the 3rd Conference on Conceptual Structure, Discourse, and Language, Boulder CO., 1997.

Lyons, J., *Linguistic Semantics: An Introduction*, Cambridge: Cambridge UP, 1995.

Ma, J.-H., *A Study of the Mandarin Chinese Verb Suffix ZHE*, Taipei: Crane Publishing. 1985.

Mei, K., "Ba Sentences in Mandarin Chinese," *Wen-shi-zhe Xuebao* 27, 1978, pp. 145-180. Taiwan University, Taipei.

Paris, M.-C., "Durational Complements and Verb Copying in Chinese," *Tsing Hua Journal of Chinese Studies New Series* 28.2, 1988, 423-439.

Pustejovsky, J., *The Generative Lexicon*, MIT Press, 1993.

Pustejovsky, J., S. Bergler, and P. Anick, "Lexical Semantic Techniques for Corpus Analysis," *Computational Linguistics* 19.2, 1993, pp. 331-358.

Ren, X., "The Post-Verbal Constituent in Chinese Passive Forms," *Journal of Chinese Linguistics* 19, 1991, 221-241.

Smith, C. S., *The Parameter of Aspect*, Kluwer Academic Publisher, 1991.

Sun, C., "Transitivity, the Ba Construction and Its History," *Journal of Chinese Linguistics* 23, 1995, pp. 159-195.

Teng, S.-H., *Chinese Synonyms Usage Dictionary*, Taipei: Crane Publishing, 1994.

Thompson, S. A., "Transitivity and the Ba Construction in Mandarin Chinese," *Journal of Chinese Linguistics* 15.1, 1973, pp. 208-221.

Tsai, M.-C., C.-R. Huang, and K.-J. Chen, "You jinyici bianyi biaozhun kan yuyi jufa zhi hudong. (From near-synonyms to the interaction between syntax and semantics)," in *Proceedings of 5th International Symposium on Chinese Languages and Linguistics* (*IsCLL 5*), 1996, pp. 167-180.

Acknowledgment

This paper is partially sponsered by National Science Council (87-2811-H-001-004). The authors would like to thank the CKIP group members for stimulating discussions on some of the verbal semantic features put forth here. We are also grateful to two anonymous reviewers for their valuable comments.

Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 75-100
©Computational Linguistics Society of R.O.C.

White Page Construction from Web Pages for Finding People on the Internet

Hsin-Hsi Chen* and Guo-Wei Bian*

Abstract

This paper proposes a method to extract proper names and their associated information from web pages for Internet/Intranet users automatically. The information extracted from World Wide Web documents includes proper nouns, E-mail addresses and home page URLs. Natural language processing techniques are employed to identify and classify proper nouns, which are usually unknown words. The information (i.e., home pages' URLs or e-mail addresses) for those proper nouns appearing in the anchor parts can be easily extracted using the associated anchor tags. For those proper nouns in the non-anchor part of a web page, different kinds of clues, such as the spelling method, adjacency principle and HTML tags, are used to relate proper nouns to their corresponding E-mail addresses and/or URLs. Based on the semantics of content and HTML tags, the extracted information is more accurate than the results obtained using traditional search engines. The results can be used to construct white pages for Internet/Intranet users or to build databases for finding people and organizations on the Internet. Such searching services are very useful for human communication and dissemination of information.

Keywords: proper name identification, information extraction, white pages, World Wide Web

1. Introduction

With the rapid growth of the Internet in recent years, the World Wide Web (WWW) has become a powerful medium for human communication and dissemination of information. Because more online information is disseminated through this giant media, the Web forms a very large knowledge resource. The explosive growth of the WWW has involved more than 10 million documents. Some search engines and information discovery systems have been introduced to help users locate relevant information. However, one

*Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. E-mail: hh_chen@csie.ntu.edu.tw

of the problems in cyberspace is that it is very difficult to know how to contact an entity, which is a concrete object that can send and receive information. For communication purposes, we usually want to know a person's or a company's E-mail address, or his/her home page URL. White pages, which are E-mail directories or URL directories in this case, can help users find such information. There are two major difficulties in building Internet white pages and people searching services. First, it is hard to set up such a white page manually because the WWW is a very large database and is created dynamically. Second, the approaches adopted by current search engines are not suitable for searching the e-mail addresses and home page URLs of people and organizations.

Current search engines only index the contents (words) of a web page with the page's URL. When a page contains many proper names, the search engine will index all the proper nouns with this page's URL. However, only one of these proper nouns or none is the owner of the page. Example 1(a) shows the appearance of a portion of a web page (http://www.ntu.edu.tw/TANet/public.html). Totally, there are 42 public universities and colleges listed in this page. The traditional search engines will index all of these 42 proper nouns with the page's URL, but none of the home pages of these proper nouns is this page.

Example 1(a). http://www.ntu.edu.tw/TANet/public.html (Browsing View)

公立大學暨獨立學院
Public University and College
國立臺灣大學 / National Taiwan University
台北市(10764)羅斯福路四段1號
1, Roosevelt Rd., Sec.4, Taipei, Taiwan, R.O.C.
Tel.:(02)3630231
Fax :(02)3627651
國立政治大學 / National Chengchi University
台北市(11623)指南路二段64號
64, Chih-Nan Rd., Sec.2, Taipei, Taiwan, R.O.C.
Tel.:(02)9393091
Fax :(02)9398043
國立清華大學 / National Tsing Hua University
新竹市(30043)光復路二段101號
101, Kuang-Fu Rd., Sec.2, Hsinchu, Taiwan, R.O.C.
Tel.:(035)715130
Fax :(035)722467
國立交通大學 / National Chiao Tung University
新竹市(30050)大學路1001號
1001, Ta-Hsueh Rd., Hsinchu, Taiwan, R.O.C.
Tel.:(035)712121
Fax :(035)721500
…………………………...

Example 1(b). http://www.ntu.edu.tw/TANet/public.html (Original HTML)

```
 <html>
 <head>
 <title>公立大學暨獨立學院 / Public University and College</title>
 <!--版權所有：國立臺灣大學對此文件保留所有權利-->
 <!--Copyright (c) 1996 National Taiwan University ALL RIGHTS RESERVED-->
 </head>    <p>
 <h2>公立大學暨獨立學院<br>Public University and College</h2>
 <ul type=square>

 <li><a href="http://www.ntu.edu.tw/">國立臺灣大學 / National Taiwan University</a>
    <ul type=disc>
          <li>台北市(10764)羅斯福路四段 1 號
          <li>1, Roosevelt Rd., Sec.4, Taipei, Taiwan, R.O.C.
          <li>Tel.:(02)3630231
          <li>Fax :(02)3627651
    </ul>
 <li><a href="http://www.nccu.edu.tw/">國立政治大學 / National Chengchi University</a>
    <ul type=disc>
          <li>台北市(11623)指南路二段 64 號
          <li>64, Chih-Nan Rd., Sec.2, Taipei, Taiwan, R.O.C.
          <li>Tel.:(02)9393091
          <li>Fax :(02)9398043
    </ul>
 <li><a href="http://www.nthu.edu.tw/">國立清華大學 / National Tsing Hua University</a>
    <ul type=disc>
          <li>新竹市(30043)光復路二段 101 號
          <li>101, Kuang-Fu Rd., Sec.2, Hsinchu, Taiwan, R.O.C.
          <li>Tel.:(035)715130
          <li>Fax :(035)722467
    </ul>
 <li><a href="http://www.nctu.edu.tw/">國立交通大學/ National Chiao Tung University</a>
    <ul type=disc>
          <li>新竹市(30050)大學路 1001 號
          <li>1001, Ta-Hsueh Rd., Hsinchu, Taiwan, R.O.C.
          <li>Tel.:(035)712121
          <li>Fax :(035)721500
    </ul>
 …………..
</html>
```

Nevertheless, the original HTML data shown in Example-1(b) shows some information about these proper nouns. Some anchor tags are provided for users to browse their home pages. For example, the URL of the home page of National Taiwan University (國立臺灣大學) is http://www.ntu.edu.tw/, which is described by the HTML anchor tag. Considering the context and HTML tags, these proper nouns and the associated anchor tags can be extracted from the web page. For example, 'National Tsing Hua University' (' 國立清華大學 ') should be related to the URL http://www.nthu.edu.tw/. On the contrary, the traditional search engines will index this school and others with the web page's URL (http://www.ntu.edu.tw/TANet/public.html). In terms of accuracy, the approach adopted by the current search engines is not suitable for such a task of finding people and organizations. Additionally, most of the anchors and the contents in a web page are not proper nouns. Such anchors and words should not be extracted and indexed for people searching.

Furthermore, some proper nouns may appear in the non-anchor part of a web page. If there are URLs of web pages and e-mail addresses on the same page, the relationships between the proper nouns and the information should be identified. Example 2 shows such a case. In contrast to the anchor part, no explicit HTML tags indicate the relationship between the proper nouns and e-mail addresses.

Example 2 - 區域網路線路維護流程

北區區域網路中心(臺大計中網路組)

……

網路組成員:

組長: 游張松 教授
E-mail: yucs@ccms.ntu.edu.tw
Tel: 3627734 ext 219

組員: 胡 湘 小姐	李光偉 先生
E-mail: giraffe@ccms.ntu.edu.tw	edward@ccms.ntu.edu.tw
Tel: 3627734 ext 241	3627734 ext 241
組員: 曾珀雯 小姐	徐信權 先生
E-mail: popo@ccms.ntu.edu.tw	kevins@ccms.ntu.edu.tw
Tel: 3627734 ext 241	3627734 ext 241

……

How to identify the proper nouns in a web page is a critical problem for building white pages. Fortunately, a very large portion of the WWW is composed of natural language documents, which can be regarded as a text corpus. Corpus analysis techniques in natural language processing [CL, 1993] can be employed to extract knowledge from the WWW. And using the semantics of the content and HTML tags, the information (URLs and e-mail addresses) can be related to proper nouns. This paper will propose a method to construct white pages for Internet/Intranet users automatically. It extracts information, including proper nouns, E-mail addresses and home page URLs, from WWW documents, and finds the relationships among these data. The problems to be tackled are as follows:

(1) Proper nouns, which are always unknown words, have to be identified and classified from a WWW corpus. Personal names and organization names are the requested entities for people finding on the Internet. Those proper nouns that denote organizations are usually hierarchical. Such relationships must be distinguished.

(2) There may be more than one proper noun, more than one E-mail address, and more

than one URL in a WWW document. Thus, we have to find a mapping from a set of E-mail addresses (or URLs) to a set of proper nouns.

The extraction method proposed in this paper was tested on the web pages in Taiwan. Section 2 introduces WWW documents and the semantics of the HTML annotation. The hierarchical nature and the related HTML tagging (1996) are discussed. Section 3 gives an overview of our white page constructor. Section 4 presents the identification algorithms for proper nouns. Here, we focus on personal names and organization names. Section 5 touches on the algorithms for mapping between proper nouns and related information. Section 6 discusses the experiments, and Section 7 offers some conclusions.

2. WWW Documents

The first step in constructing white pages is to find out where proper nouns, E-mail addresses and URLs are located in WWW documents. Web documents are different from a traditional text corpus in that they are HTML (HyperText Markup Language) files. The tagging information provides some clues, but it also introduces some noise. How to use the information is a very important issue in applications on the Internet, e.g., cross-language information retrieval [Bian and Chen, 1997]. In plain text, each sentence always has a sentence terminator, such as a full stop, question mark or exclamation mark. These symbols split each document into several processing units. In HTML files, these punctuation marks do not always appear. Quasi-sentences are defined according to some HTML tags shown below:

●Title (TITLE)
●Headings (H1, H2, ⋯ , H6)
●Address (ADDRESS)
●Unordered Lists (UL, LI)
●Ordered Lists (OL, LI)
●Definition Lists (DL, DT, DD)
●Tables (TABLE, TD, TH, TR)

Furthermore, some punctuation symbols like '|' and ':' have the same effects. In contrast to the above sentence delimiters, the font style elements may introduce noise. Bold (B), italic (I), superscripts (SUP), subscripts (SUB) and font (FONT) can be used to emphasize some points in texts. However, these elements produce many unknown words because a word is split into several parts by HTML tags. Example 3 illustrates the word 'Font' associated with various font style tags. Thus, these tags should be treated as meta-information and hidden from processing.

Example 3.

```
<I><B><FONT SIZE=+2><FONT COLOR="#FF0000">F</FONT>
<FONT COLOR="#0000FF">o</FONT>
<FONT COLOR="#FF8000">n</FONT>t</FONT></B></I>
```

Links denoted by anchors (A) in WWW documents are possible sources of proper nouns and related information. The WWW documents shown in Appendix A shows their typical features. The first example is the home page of National Taiwan University (NTU, http://www.ntu.edu.tw/). The entity that we are interested in is 'National Taiwan University' (' 國立台灣大學 '), which is an organization name and is shown in the title area. The second example (http://www.ntu.edu.tw/NTULink/) follows from 'NTU Link' on the NTU home page. An underline shows a link to other home pages in the web page. The interesting entities are Office of Academic Affairs (' 教務處 '), Office of Student Affairs (' 學務處 '), and Office of Business Affairs (' 總務處 '); University Library (' 圖書館 '); Computer and Information Network Center (' 計算機及資訊網路中心 '); Population Studies Center (' 人口研究中心 '). Those units that do not have any links are not considered. For example, the home pages for Accounting Office (' 會計室 ') and Military Instructors' Office (' 軍訓室 ') have not yet been constructed now, so that they are not listed in the final white pages. Following the link for 'Colleges, Schools, Departments, Graduate Institutes and Affiliated Organizations', we can retrieve more information. All these units form a hierarchical structure in National Taiwan University.

A link in the HTML file may be represented as follows:

```
<a href="argument"> text </a>
```

When "text" is a proper noun, its home page URL can be described by "argument". Consider an example on the 'NTU Link' web page. The link to 'Office of the Dean of Academic Affairs' (' 教務處 ') is shown below:

```
<a href="/Campus/announce/index.html#academic"> 教務處
/ Office of the Dean of Academic Affairs</a>
```

If the proper noun and its URL are put into white pages directly, this entry may be ambiguous. This is because many universities have similar organizations. Therefore we should keep the hierarchical path of the web page to disambiguate the meaning of a proper noun. Further, the relative URLs have to be changed into absolute ones to keep all of the URL information. Because the URL associated with the link 'Office of the Dean of Academic Affairs' (教務處) is a relative URL

(/Campus/announce/index.html#academic) and the web page's URL is http://www.ntu.edu.tw/NTULink/, the absolute URL of this organization is represented as http://www.ntu.edu.tw/Campus/announce/index.html#academic. In addition, the host name (www.ntu.edu.tw) in the hierarchical path of this URL shows that this organization is part of National Taiwan University. The complete organization name will be 'Office of the Dean of Academic Affairs in National Taiwan University' (' 國立台灣大學 教務處 '). Therefore, similar organizations and personal names can be disambiguated with the host names of their absolute URLs to find their home pages' URLs on the global Internet.

Besides the linking anchor field, proper nouns may appear in other portions of a WWW document. Dealing with these objects is more complex because no explicit HTML tags indicate the URLs of these objects. An additional algorithm is needed to associate URLs and E-mail addresses with suitable proper nouns. Different kinds of clues, such as the spelling method, adjacency principle and HTML tags (e.g., title, headings, address, and font style elements), are employed.

3. System Overview

We periodically collect web pages from the Internet/Intranet using a spider. The white page constructor first analyzes these HTML files. Basic processing units (sentences or quasi-sentences) and HTML meta-information are gathered. Because a Chinese sentence (or quasi sentence) is composed of a sequence of characters without word boundaries [Chen and Lee, 1996], a Chinese segmentation system identifies the word tokens. Then, a proper noun identification system (see Section 4) extracts personal names and organization names. During processing, the information in the anchor parts is placed in the anchor set (AS). Other information, i.e., that appearing in non-anchor parts, is placed in one of the content sets (CSes) which correspond to different types of information. In the current implementation, there are three content sets: CS_Proper-Noun, CS_E-Mail and CS_HTTP. They record proper nouns, E-mail addresses and URLs, respectively. For the anchor set, the remaining task is simple. We just relate the proper noun found in an anchor to the corresponding URL or E-mail address. For the content sets, a mapping algorithm (see Section 5) is proposed to associates URLs and/or E-mail addresses with a suitable proper noun. Algorithm 1 shows the information extraction part of the white page constructor.

Algorithm 1. Information Extraction

```
Input:    An HTML file or a plain text with its URL (URL_1)
Output:   An anchor set (AS) and three content sets (CSs)
Method:   1. [HTML Parser]
             Identify sentence boundary and collect those HTML tags that
             are useful for information mapping.

          2. [Chinese Segmentation System]
             For each processing unit (a sentence or a quasi-sentence),
             identify the word boundary.

          3. [Identification of Proper Nouns]
             Identify and classify proper nouns in the text.

          4. For each proper noun (PN)
             {
             4.1 [Analyze the <Title> tag: <TITLE>Title_Text</TITLE>]
                 if PN tagged with the HTML tag <Title>,
                    add the tuple (PN, URL_1) to the Anchor Set (AS)
             4.2 [Extract the Anchor Information]
                 if PN tagged with the HTML tag <A>
                      (<a href=" protocol://host/path ">Text</a>),
                    add the tuple (PN, protocol://host/path) to the Anchor
                      Set (AS)

             4.3 [Extract the Content Information]
                 if PN is in the non-anchor part (content)
                       add PN to CS_Proper-Noun with the following
                           attributes:
                           the position information of token (token_no) and
                           the associated HTML meta information (<TITLE>
                           <Hn>, <Address>, <Bold>, <Font> and <Italic>)
             }
          5. Extract different types of information with the position
                information of token (token_no), and add to the
                corresponding Content Sets (CS_E-Mail and CS_HTTP)

          6. End
```

4. Identification of Proper Nouns

Proper nouns that are not collected in lexicons are major unknown words in natural language texts. Several methods [Boguraev and Pustejovsky, 1996; Mani, *et al*., 1993; McDonald, 1993; Paik, *et al*., 1993] have been proposed to identify English proper nouns. For research related to Chinese, Chang *et al*. [1992] and Wang *et al*. [1992] touched on Chinese personal names; Sproat *et al*. [1994] considered Chinese personal names and transliterations of foreign words; Chen and Lee [1996] identified Chinese personal names, Chinese transliterated personal names and organization names. The name identification module is based on our previous design. The methods are described below.

4.1 Identification of Personal names

A Chinese personal name is composed of surname and given name parts. Most Chinese surnames are single characters (model (a)), and some rare ones have two characters (model (b)). A married woman may place her husband's surname before her surname (model (c)). Thus there are three possible types of surnames, i.e., single character, two characters and two surnames together. Most names have two characters, and some rare ones are single characters. Theoretically, every character can be considered as a names rather than a fixed set. Thus, the length of Chinese personal names ranges from 2 to 6 characters. The baseline models for identification are shown as follows.

Model (a) Single character surname:

$$(1)\ \frac{\# C_1}{\& C_1} \times \frac{\# C_2}{\& C_2} \times \frac{\# C_3}{\& C_3} > Threshold1$$

$$(2)\ \frac{\# C_1}{\& C_1} > Threshold2 \text{ and } \frac{\# C_2}{\& C_2} \times \frac{\# C_3}{\& C_3} > Threshold3$$

Model (b) Two characters surname:

(3) $C_{11}C_{12}$ is two-character surname and $\frac{\# C_2}{\& C_2} \times \frac{\# C_3}{\& C_3} > Threshold4$

Model (c) Two surnames together:

$$(4)\ \frac{\#C_{11}}{\&C_{11}}\times\frac{\#C_{12}}{\&C_{12}}\times\frac{\#C_{2}}{\&C_{2}}\times\frac{\#C_{3}}{\&C_{3}} > Threshold5$$

$$(5)\ \frac{\#C_{11}}{\&C_{11}}\times\frac{\#C_{12}}{\&C_{12}} > Threshold6 \text{ and} \frac{\#C_{2}}{\&C_{2}}\times\frac{\#C_{3}}{\&C_{3}} > Threshold7$$

where C_1, C_{11}, and C_{12} are the characters forming a surname,
C_2 and C_3 are the characters which are considered as names,
$\# C_i$ is the frequency of C_i being a surname or a name,
$\& C_i$ is the frequency of C_i being contained in the other words.

For different types of surnames, the different models are adopted. Because the two-character surnames are always indicated as surnames, Model (b) neglects the score of the surname part. Models (a) and (c) have two score functions. They solve the problem of very high scores of surnames. The above three models can be extended to single-character names by ignoring the last character C_3 in each formula for training and testing. When a candidate cannot pass the thresholds, its last character is cut off and the remaining string is tried again. The frequencies of characters being surnames or names are trained from a large-scale Chinese name corpus of 219,738 Chinese personal names and 661,512 characters. The frequencies of characters being other words are trained from an NTU balanced corpus to compute the variation of characters. In total, this corpus has 113,647 words and 191,173 characters. Thresholds are trained using the Chinese name corpus. We calculate the scores of all Chinese personal names in the corpus using the above formulas. The scores for each formula are sorted, and then the one that is less than 99% of the personal names is considered to be a threshold for this formula. That is, 99% of the training data can pass the threshold.

Chinese personal names are not always composed of single characters. For example, the name part ' 聰明 ' (Cong-ming) of the sentence ' 陳聰明 醫術 非常 高明 ' (Chen Cong-ming yishu feichang gauming; Chen Cong-ming has find command of the medical art) is a word. How to tell that a word is a content word or a name is indispensable. Mutual information [Church and Hanks, 1990], which provides a measure of word association, is employed to differentiate between a name and a content word. We check the string that can serve as a name or a content word with its surrounding words. When they have a strong relationship, it has high probability of being a content word rather than a name. In the example ' 陳 家世 清白，絕 不會 犯法… ' (Chen jashi qingbai jue buhui fanfa … ; Chen has a clean family background and will never violate the law …), the two words ' 家世 ' (jashi) and ' 清白 ' (qingbai) have high mutual information, so that ' 陳 家世 ' (Chen jashi) is not a personal name in this example. Three newspaper corpora (total size about 2.6 million words) are used to train the word association.

Punctuation marks play an important role in identification. Personal names usually appear at the head or the tail of a sentence. A candidate is given an extra bonus when it is found in one of these two places. Gender has a special role in Chinese personal names. A married woman may place her husband's surname before her surname. That forms the personal name of model (c). Gender information helps us to disambiguate the type of personal name.

The last clue is the paragraph information. A personal name may appear more than once in a paragraph. This phenomenon is useful during identification. We use a cache to store identified candidates and reset the cache before next paragraph is processed. Consider the examples ' 焦仁和 表示 … ' (Jiao Renhe biaoshi … ; Jiao Renhe expressed …) and ' 焦仁和 秘書長 … ' (Jiao Renhe mishuzhang ...; Jiao Renhe Chief Secretary ...). Two candidates ' 焦仁 ' (Jiao Ren) and ' 焦仁和 ' (Jiao Renhe) are proposed and stored in the cache, but the personal name is finally identified as ' 焦仁和 ' (Jiao Renhe). For details, the reader is referred to a previous paper [Chen and Lee, 1996].

4.2 Organization Names

The structure of organization names is more complex than that of personal names. Basically, a complete organization name can be divided into two parts, i.e., name and keyword. Many words can serve as names, but only some fixed words can be regarded as keywords. Thus, keywords are important clue used to extract organization names. However, there are still several difficult problems. First, a keyword is usually a common content word. It is not easy to differentiate between a keyword and a content word. This problem results in ambiguities in POS tagging and word sense. Second, a keyword may appear in an abbreviated form. Third, a keyword may be omitted completely. Fourth, some organization names are very long, so it is hard to decide on the left boundary. The following examples illustrate these problems.

(1) Ambiguity of keywords:

(1.1) Ambiguity of word senses and POS tagging:

學會 (xuehui; Association or Learn)

(1.2) Ambiguity of POS tagging (verb or noun):

調查中心 (iaocha zhongxin; center of investigation), 研究中心 (yanjiu zhongxin; center of research), 開發公司 (kaifa kongsi; company of development), 開發中心 (kaifa zhongxin; center of development), 發展協會 (fazhan xiehui; development association), 規劃小組 (guihua xiaozu; planning group), 研習社 (yanxishe; research club), 評論社 (pinglunshe; discussion club), 發明社 (famingshe; invention club), 聯誼會 (lianyihui; social gathering)

(2) Abbreviated keywords:

投顧 (' 投資顧問公司 '): tougu (touziguwengongsi); Security Investment Consulting
護專 (' 護理專科學校 '): huzhuan (hulizhuankexuexiao); college of nursing
專校 (' 專科學校 '): zhuanxiao (zhuankexuexiao); college for professional training
工專 (' 工業專科學校 ') gongzhuan (gongyezhuankexuexiao); college of technology
商專 : shangzhuan; college of commerce
藝專 : yizhuan; college of arts
實小 (' 實驗小學 '): shixiao (shiyanxiaoxue); experimental primary school

(3) Keyword omitted:

宏碁 (hongji; Acer), 友訊科技 (youxunkeji, D-Link Tech.),
友力資訊 (youlizixun; Ulead Inc.)

(4) Long organization names:

國立台灣工業技術學院 (guoli taiwan gongye jishuxueyuan; National Taiwan Institute of Technology), 國家地震工程研究中心 (guojia dizhen gongcheng yanjiu zhongxin; National Center for Research on Earthquake Engineering),
實踐設計管理學院 (shijian sheji guanli xueyuan; Shih Chien College of Design and Management), 台北市大安區萬芳社區發展協會 (taibei shi daan qu wanfang shequ fazhan xiehui; Taipei Daan District Wanfang Community Development Association)

Our previous work [Chen and Lee, 1996] only touched on the fourth problem. Keywords, which are good indicators, play a role similar to that of surnames. They show not only the possibility of an occurrence of an organization name, but also its right boundary. A prefix is a good marker for a possible left boundary, for example, ' 國立 ' (National), ' 省立 ' (Provincial), and ' 私立 ' (Private), and so on. The name part of an organization may consist of single characters or words. Parts of speech, such as transitive verbs, adjectives, numerals and classifiers, are also useful for determining the left boundary. The name part of an organization cannot cross these critical parts of speech. For example, ' 公司 ' (company) in ' 三家公司… ' (three company …) is not a keyword due to the critical parts of speech. Because a tagger is not involved before identification, the part of speech of a word is determined wholly based on its lexical probability.

Although our previous experiment has shown that these critical parts of speech are useful in determining the left boundary of an organization name in a newspaper text, the ambiguity of parts-of-speech (as verb or noun) decreases the performance for the specific task - identification of organization names in an unrestricted domain. For example, the identification system will miss or give an incorrect left boundary for organization names containing ' 調查 ' (diaocha; investigation), ' 研究 ' (yanjiu; research), ' 開發 ' (kaifa; development), ' 發展 ' (fazhan; development), ' 規劃 ' (guihua; plan), ' 研習 ' (yanxi; study), ' 評論 ' (pinglun; critique), ' 設計 ' (sheji; design), ' 管理 ' (guanli; management), ' 發明 ' (faming; invention), and so on. To resolve this problem in proper noun extraction, a refined method is proposed to deal with such organization names. The experiments described in Section 6 will illustrate the performance of the baseline and refined methods.

5. A Mapping Algorithm

Identified proper nouns may appear in the anchor parts or the non-anchor parts of HTML files. For proper nouns in anchor parts, the anchor tags indicate their home pages' URLs or e-mail addresses explicitly. Consider the example "<a href="http://www.ntu.edu.tw/"> 國立臺灣大學 / National Taiwan University </a>". The URL of the home page of National Taiwan University (國立臺灣大學) is http://www.ntu.edu.tw/ as described by the HTML anchor tag. Based on the HTML tags, the information about these proper nouns attributed by anchor tags can be extracted easily from web pages.

For proper nouns appearing in non-anchor parts, a more complicated procedure is employed. Because the relationships between the proper nouns and the corresponding information are not specified explicitly, a mapping scheme can be used to associate URLs and e-mail addresses with suitable proper nouns. The following shows an example.

Example 4 - 各系所網路管理人 (the network manager of each department)
(http://www.ntu.edu.tw/NTUCC/NetManager.html)

各系所網路事務之管理負責人

DEPNAME	ROUTEMAN	ROUTETEL	ROUTEOFF	EMAIL
電機工程學系、所	王凌霄	3212-4 ext 234	電機工程學系 234 室	
材料研究所	曾德玉	3638912	工綜館 625 室	
化工系、所	吳名弘	2185	化工系機算機室	
資訊系、所	黃育銘	3625336-221	資訊新館 2F221	root@csman.csie.ntu.edu.tw
……..				
地理系、所	蔡博文	2147	地理系管二樓	tsaibw@ccms.ntu.edu.tw
地質系、所	蕭銘璽	2341 ext. 13	地質系 313 室	r2204204@sun03.gl.ntu.edu.tw
動物系、所	丘台生	2128	漁科館 401 室	tschiu@ccms.ntu.edu.tw
……				

Algorithm 2 illustrates the mapping between URLs (and/or E-mail addresses) and proper nouns. A score function that considers the spelling method, adjacency principle and HTML tags is used to determine the relationships among proper nouns and related information.

The ranking function is defined as follows:

$$\mathit{Score}(\text{Info}, \text{PN}) =$$

$$\left(\frac{\text{HTML\_SCORE(PN)} + 1}{\text{abs(Info.token\_no - PN.token\_no)}} + \frac{\text{Title(PN)}}{\text{Total\_tokens - Info.token\_no} + 1} \right)$$
$$+ \text{Pinyin\_Similarity(PN, Info)} * \text{E - mail(Info)} * \text{Weight}$$

HTML_SCORE(PN) =

$$\mathit{Title}(PN) + \mathit{Heading}(PN) + \mathit{Address}(PN) + \mathit{Bold}(PN) + \mathit{Font}(PN) + \mathit{Italic}(PN)$$

where Info is a URL or an e-mail address,

PN is a proper noun,

Info.token_no and PN.token_no are the positions of the specific tokens,

Total_tokens is the total number of tokens in the file,

Title(), Heading(), Address(), Bold(), Font(), Italic() and E-mail() are the Boolean functions,

Pinyin_Similarity(PN, Info) is defined below and used to measure the similarity between PN and Info under the criteria of Pinyin,

Weight is used to measure the importance of Pinyin similarity.

```
Algorithm 2. Information Mapping
Input:      Three Content Sets (CSs)
            A Threshold and a Window_Size of context
Output:     A Mapping Set (MS)
Function:   Mapping CS_E-mail (CS_HTTP) to CS_Proper-Name
Method:     1. Set MS to be an empty set.
            2. For each CS information set (i.e., CS_E-mail and CS_HTTP)
               {  /* the mapping between CS and CS_Proper-Noun may be
                     Many-to-one. */
                  copy CS_Proper-Noun to CD
                  for each entry Info in CS
                  {  PN is an entry whose offset from Info is less than
                        Window_Size, and Score(Info, PN) is the maximum
                        in CD.  If many entries have the same maximum
                        value, the entry appearing before Info is chosen.
                     if Score(Info, PN) > Threshold
                     {    add (Info, PN) into MS
                     }
                  }
               }
           3. End
```

The *Score* function combines the following heuristic rules:

(1) **Spelling Method**. If the extracted information (Info) is an E-mail address, the similarity between Info and the proper noun (PN) is considered. Because the user-id in an E-mail address is often transliterated from a Chinese name, this heuristic rule is preferred over other cues, and we assign it a larger weight. The Pinyin system [Lu, 1995] is adopted to transliterate Chinese names. For robustness, the Pinyin similarity is defined as follows:

$$\mathit{Pinyin\_Similarity}(\text{PN, E-mail}) = \frac{\text{\# of letters in user - id that match the pinyin transliteration of PN}}{\text{total \# of letters in the user - id of the E - mail address}}$$

where PN is a proper noun, and E-mail is an e-mail address.

For example, the Pinyin transliteration of " 邊國維 " is "Bian Guo Wei". The similarities between the following e-mail addresses and this personal name are:

$$\text{Pinyin\_Similarity(邊國維 , gwbian@nlg.csie.ntu.edu.tw)} = \frac{6}{6} = 1$$

$$\text{Pinyin\_Similarity(邊國維 , arthur\_bian96@nlg.csie.ntu.edu.tw)} = \frac{4}{10} = 0.4$$

$$\text{Pinyin\_Similarity(邊國維 , arthur@nlg.csie.ntu.edu.tw)} = \frac{0}{6} = 0$$

(2) **Adjacency Principle.** Proper nouns and the related information are often close to each other. The distance between Info and PN is measured in terms of the number of intervening tokens. Recall that we assign each object a unique token number. Closer pairs have larger scores. Additionally, a proper name appearing in the title of a web page (tagged with <Title>) will be treated close to the rear of a web page.

(3) **HTML Tags.** Proper nouns (PNs) that appear in Title (<Title>), Heading (<Hn> ··· </Hn>) or Address, or are described by the font style (Bold, Italic and Font tag elements) are given larger weights than other normal proper nouns.

6. Experiments

In our initial experiments, a total of 703 web pages were collected from the NTU Web (http://www.ntu.edu.tw/). A person identified the personal names and organization names in these web pages and associated them with the URLs and the e-mail addresses if possible. Then, the collected answers were classified into an anchor set and a content (non-anchor) set.

The results of identification using the proposed system were checked against human results. The window size (Window_Size) of context was 6, and the score threshold (Threshold) was 0.2 for the mapping algorithm. The threshold was greater than the inverse of the window size. It was used to filter out proper nouns that were near the window boundary but were not described by any HTML tags.

Table 1 shows the results of identification in both sets and the mapping result in the content set. In Table 1(a) and 1(b), the number of personal names and the number of organization names identified by humans are listed in column 2. Columns 3 and 4 show the identification results of proper nouns using our system and the correct results. The precision and the recall are defined as follows.

$$\text{Precision} = \frac{\#\text{ of items identified correctly by program}}{\#\text{ of items identified by program}}$$

Table 1. *The Results of Identification and Information Mapping*

(a) Identification of Proper Nouns in the Anchor Set

Anchor Set	# of items in the web pages of NTU	# of items identified by program	# of items identified correctly by program	Precision	Recall
Personal name	255	228	189	82.89%	74.12%
Organization Name	746	611	213	34.86%	28.55%

(b) Identification of Proper Nouns in the Content Set

Content Set	# of items in the web pages of NTU	# of items identified by program	# of items identified correctly by program	Precision	Recall
Personal name	1732	3343	1470	43.97%	84.87%
Organization Name	3029	2272	503	22.14%	16.61%

(c) Identification of Organization Names Using Refined Method

Organization Name	# of items in the web pages of NTU	# of items identified by program	# of items identified correctly by program	Precision	Recall
Anchor Set	746	856	558	65.19%	74.80%
Content Set	3029	3392	2082	61.38%	68.74%

(d) The Mapping Result in the Content Set

Content Set Mapping	# of items extracted by program	# of items mapped correctly by program	# of items mapped incorrectly by program	Accuracy
E-mail	64	18	5	78.26%
HTTP	16	1	0	100%

$$\text{Recall} = \frac{\text{\# of items identified correctly by program}}{\text{\# of items in the web pages}}$$

In the anchor part, there were 6,204 linking items. Of these, the numbers of personal names and organization names were 255 and 746, respectively. That is, 83.87% of the anchors were irrelevant and should be screened out for the task of finding people. The precision and the recall rates were 82.89% and 74.12% for the identification of personal names, respectively.

However, the precision and the recall rates for the identification of organization names were much lower than those obtained in our previous work. The major errors resulted from the strategy discussed in Section 4.2, i.e., "parts of speech such as transitive verbs, adjectives, numerals and classifiers are also useful to determine the left boundary, and the name part of an organization cannot cross these critical parts of speech." Many of the organization names may contain ' 調查 ' (diaocha; investigation), ' 研究 ' (yanjiu; research), ' 開發 ' (kaifa; development), ' 發展 ' (fazhan; development), ' 規劃 ' (guihua; plan), ' 研習 ' (yanxi; study), ' 評論 ' (pinglun; critique), ' 設計 ' (sheji; design), ' 管理 ' (guanli; management), ' 發明 ' (faming; invention), and so on. All of these words can be nouns or transitive verbs. The identification system misses or gives the incorrect left boundary for such an organization name. The following examples illustrate this problem.

> 公共政策研討學會 (gonggong zhengce yantao xuehui; Public Policy Workshop Association), 科學管理學會 (kexue guanli xuehui; Science Management Association), 中央研究院調查研究工作室 (zhongyang yanjiuyuan diaocha yanjiu kongzuoshi; Office of Survey Research at Academia Sinica), 電影研究社 (dianyin yanjiushe; Movie Club), 機車研習社 (jiche yanxishe; Motorcycle Club), 女青年聯誼會 (nuqingnian lianyihui; Youth Women's Christian Association), 舞台設計工作室 (wutaisheji gongzuoshi; Studio of Stage Design), 台北市大安區萬芳社區發展協會 (taibei shi daan shequ fazhan xiehui; Taipei Daan District Wanfang Community Development Association), 實踐設計管理學院 (shijian sheji guanli xueyuan; Shih Chien College of Design and Management), 台北影視開發公司 (taibei yingshi kaifa gongsi; Taipei Movie, Video and Television Development Company), 臺灣大學推廣教育中心 (taiwan daxue tuiguang jiaoyu zhongxin; Center of Extended Education, National Taiwan University), 水產試驗所 (shuichan shiyansuo; Fishery Research Institute), 台大校園規劃小組 (taida xiaoyuan guihua xiaozu; Campus Planning Group, National Taiwan University), 領導公關公司 (lingdao gongguan gongsi; Lingdao Public Relation Company), 編輯委員會 (bianji weiyuanhui; Editing Committee), 調查委員會 (diaocha weiyuanhui; Investigation Committee), 污泥處置研究所 (wunichuzhi yanjiusuo; Mud Disposal Research Institute), 交大應用藝術研究所 (jiaoda yingyong yishu yanjiusuo; Institute of Applied Arts, National Chiao Tung University), 生物技術開發中心 (shengwu jishu kaifa zhongxin; Development Center for Biotechnology), 中華民國視訊發展協會 (zhonghua shixun fazhan xiehui; Telecommunication Development Association of Republic of China), 農業試驗所 (nongye shiyansuo; Agriculture Research Institute), 開拓文教基金會 (kaituo wenjiaojijinghui; Kaituo Cultural and Educational Foundation), 國際翻譯社 (guoji fanyishe; International Translation Agency).

To resolve this problem, a refined method was used to allow these words to serve as the name parts of organization names. The performance of the refined method is shown in Table 1(c). With this heuristic rule, the precision was 65.15% and the recall was

74.79% in the anchor part. Appendix B presents some extracted examples in the anchor part. The refined method achieved a precision rate of 61.38% and a recall rate of 68.74% for the content part.

In the content part of the 703 web pages, there were 1,732 proper names and 3,029 organization names. Only one of these proper nouns or none was the owner of the web page. That is, at least 85.23% of these names were unrelated to the owners of the web pages. Totally, 64 E-mail addresses and 16 HTTP URLs were extracted in the non-anchor part. Because the patterns of the E-mail addresses and HTTP URLs were well-formed, all of them were found. These addresses and URLs were related to none or one of 6,735 proper nouns (3,343 personal names and 3,392 organization names). With the mapping heuristics, 18 E-mail addresses were assigned to the correct personal names or organization names; 5 E-mail addresses were assigned incorrectly; and the others were not assigned. The mapping algorithm achieved an accuracy rate of 78.26%. We found that the Pinyin spelling similarity provided a very good criterion to relate the E-mail addresses to the proper nouns, even when they were not the nearest pairs. Some experimental data and results are shown in Appendix C.

Table 2 summarizes the overall results of information extraction for proper nouns. 97.52% of the information was extracted from the anchor set. The number of home page URLs and E-mail addresses extracted in the content part was much smaller than that in the anchor part. This reflects the characteristics of web pages. When designing web pages, people often include URLs and E-mail addresses within the linking anchors for users' navigation instead of giving the information in the content. Because the HTML anchor tags explicitly give the information about the linking text, the overall performance will depend on the identification of proper nouns in the anchors of the web pages.

Table 2. *The Overall Results of Information Extraction for People Finding*

Information Extraction	# (no. of correct items)	% (Percentage)
From Anchor Set	747 (189 for people and 558 for organizations)	97.52%
From Content Set	19	2.48%
Total	766	100%

The major errors resulted from conjunctions and compounds in the organization names. For complex proper names, the correct boundaries were not determined in the identification task. Some examples are shown in the following. In the string ' 台大建築與城鄉研究所 ' (Taida Jianzhu yu Chengxiang yanjiusuo; Graduate Institute of Building and Planning), an organization name ' 城鄉研究所 ' (Chengxiang yanjiusuo) was identified with an incorrect left boundary because of the conjunction ' 與 ' (yu; and).

```
<A href="http://www.bp.ntu.edu.tw/">台大建築與城鄉研究所 / Graduate Institute of Building and Planning </A>
        Oname: 城鄉研究所
<a href="http://jojo.ntu.edu.tw/TANet/public.html">公立大學暨獨立學院 / Public University and College</a>
            Oname: 公立大學
<a href="http://jojo.ntu.edu.tw/TANet/public.html">公立大學暨獨立學院 / Public University and College</a>
            Oname: 獨立學院
<a href="http://linux1.cgu.edu.tw/">長庚醫學暨工程學院 / Chang Gung College of Medicine and Technology</a>
            Oname: 工程學院
<a href="http://jojo.ntu.edu.tw/TANet/edu.html">教育網路中心 / Educational Network Center</a>
            Oname: 網路中心
<a href="http://www.hcht.edu.tw/">華梵人文科技學院 / Huafan College of Humanities and Technolgy</a>
            Oname: 科技學院
```

In the content part of the web pages, the system produced some incorrect personal names. For example, the personal name '魏晉南' (Wei Jin Nan) was incorrectly identified in the famous dynasty '魏晉南北朝' (Wei Jin Northern and Southern Dynasties), because '魏' (Wei) is a frequently-used surname and '晉南' (Jin Nan) is like a name. Further, some famous historical books and names of years were very similar to the personal names. On the other hand, some famous ancient personal names could not be identified, because the name parts of these names were rarely used in the training corpus of contemporary personal names. In addition, nicknames and some transliterated names were missed. This is because a nickname does not lead with a surname, and most of the characters used in Japanese names are different from those in transliterated English names. Some examples are listed below.

(Incorrect Identification)

- Famous dynasty:
  魏晉南 (魏晉南北朝 ; Wei Jin Northern and Southern Dynasties),
  魏晉隋 (Wei Jin & Sui Dynasties), 魏晉 (Wei & Jin Dynasties),
  隋唐 (Sui & Tang Dynasties)
- History books:
  史傳 (Shizhuan), 白書 (Baishu), 古史 (Gushi; Ancient History)
- Year: 丁巳 (Ding Si)
- Transliterated Japanese Names: 山根 (山根幸夫 ; Yamane Yukio)

(Miss)

- Ancient People:
  司馬遷 (simaqian), 顧炎武 (guyanwu), 劉知幾 (liuzhiji), 胡適 (Hushi), 逯耀東 (luyaodong), 郭沫若 (guomuruo)
- Nicknames:
  小安安 (xiao anan), 小桂子 (xiao guizi), 阿勳 (a xun), 阿賢 (a xian),
  潔潔 (jiejie), 雄雄 (xiongxiong)

●Transliterated Japanese Names:
雄川一郎 (Okawa Ichiro), 小林直樹 (Kobayashi Naoki)

To increase the coverage of the dictionary can reduce the error rates in name recognition. For example, famous dynasties, history books, and famous ancient personal names should be added to the lexicon. Furthermore, nicknames and transliterated names (e.g., Japanese names) should be investigated further.

7. Concluding Remarks

This paper has proposed a computer-aided information extraction method to construct white pages for Internet/Intranet users or to build databases for finding people and organizations on the Internet. The traditional approach used by current search engines indexes proper nouns with incorrect URLs of web pages in the task of finding people and organizations. In our system, proper nouns are identified using some heuristic rules and the corpus-based analysis method of natural language processing. Considering the semantics of content and HTML tags, these proper nouns and their related information are extracted from web pages. Using identification of proper nouns, the number of indexing terms on a web page using the proposed method is smaller than that using search engines. Finding people and organizations in the database of the extracted results is more precise than in the current search engines. The results here show that much interesting information can be automatically extracted from the WWW. However, complete identification of conjunctions and compounds in organization names needs further investigation. Furthermore, other types of information, e.g., addresses, phone numbers, and so on, will be considered in the future.

References

Bian, G.W. and Chen, H.H. "An MT Meta-Server for Information Retrieval on WWW", *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, Palo Alto, California, USA, March, 1997, pp.10-16.

Boguraev, B. and Pustejovsky, J. *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA, USA., 1996.

Chang, J.S., et al. "Large-Corpus-Based Methods for Chinese Personal Name Recognition", *Journal of Chinese Information Processing* 6.3 (1992): pp. 7-15.

Chen, H.H and Lee, J.C. "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 15th International Conference on Computational Linguistics*, 1996, pp. 222-229.

Church, K.W. and Hanks, P. "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics* 16.1 (1990): pp. 22-29.

CL, "Special Issues on Using Large Corpora," *Computational Linguistics* 19. 1-2, 1993.

Davis, M.W. and Ogden, W.C. "Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 9-17.

Etzioni , "Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web", URL ftp://ftp.cs.washington.edu/pub/etzioni/softbots/a96.ps.gz, *Proceeding of AAAI-96*, 1996.

Gachot, D.A.; Lange, E. and Yang, J. "The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Multilingual Information Retrieval." *Proceedings of Workshop on Cross-Linguistic Information Retrieval,* 1996, pp. 44-54.

Hayashi, Y.; Kikui, G. and Susaki, S. "TITAN: A Cross-linguistic Search Engine for the WWW." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 58-65.

HTML, *HyperText Markup Language*, URL http://www.w3.org/pub/WWW/Markup, 1996.

Lu, Suping, "A Study on the Chinese Romanization Standard in Libraries," *Cataloging and Classification Quarterly* 21 (1995):81-97.

Mani, I., et al. "Identifying Unknown Proper Names in Newswire Text," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 44-54.

McDonald, D. "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 32-43.

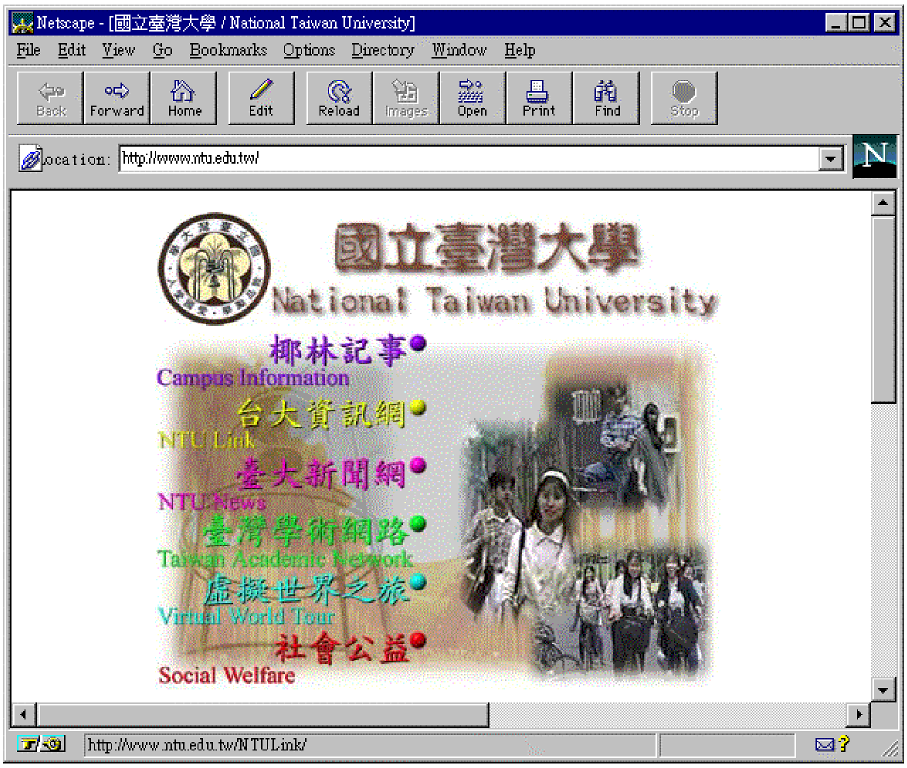
Paik, W., et al."Categorization and Standardizing Proper Nouns for Efficient Information Retrieval," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 154-160.

Sproat, R., et al."A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", *Proceeding of 32nd Annual Meeting of ACL, New Mexico*, 1994, pp. 66-73.

Wang, L.J; Li, W.C. and Chang, C.H. "Recognizing Unregistered Names for Mandarin Word Identification", *Proceeding of 14th COLING*, Nantes, 1992, pp. 1239-1243.

Appendix A. Hierarchical Features of Home Pages

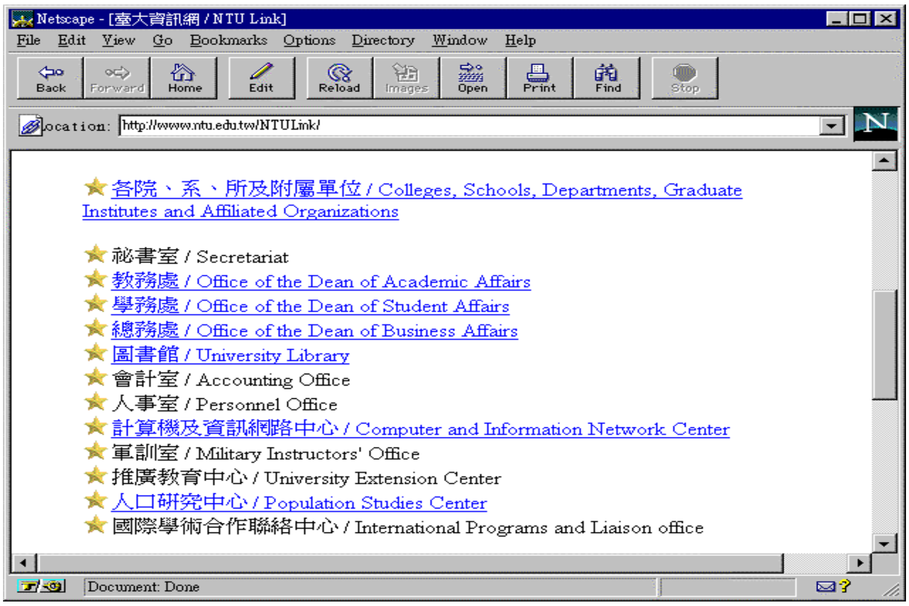
(1) Home page of National Taiwan University


Netscape - [國立臺灣大學 / National Taiwan University]

Location: http://www.ntu.edu.tw/

http://www.ntu.edu.tw/NTULink/

(2) Home Page from NTU Link


Netscape - [臺大資訊網 / NTU Link]

Location: http://www.ntu.edu.tw/NTULink/

File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Edit Reload Images Open Print Find Stop

★ 各院、系、所及附屬單位 / Colleges, Schools, Departments, Graduate Institutes and Affiliated Organizations

★ 祕書室 / Secretariat
★ 教務處 / Office of the Dean of Academic Affairs
★ 學務處 / Office of the Dean of Student Affairs
★ 總務處 / Office of the Dean of Business Affairs
★ 圖書館 / University Library
★ 會計室 / Accounting Office
★ 人事室 / Personnel Office
★ 計算機及資訊網路中心 / Computer and Information Network Center
★ 軍訓室 / Military Instructors' Office
★ 推廣教育中心 / University Extension Center
★ 人口研究中心 / Population Studies Center
★ 國際學術合作聯絡中心 / International Programs and Liaison office

Document: Done

Appendix B. Some Experimental Results in the Anchor Part

In the following, Oname and Pname denote the extracted organization names and personal names, respectively.

[Organization-School (Oname)]

```
<a href="http://www.ntu.edu.tw/">國立臺灣大學 / National Taiwan University</a>             Oname: 國立臺灣大學
<a href="http://www.nccu.edu.tw/">國立政治大學 / National Chengchi University</a>           Oname: 國立政治大學
<a href="http://www.nthu.edu.tw/">國立清華大學 / National Tsing Hua University</a>          Oname: 國立清華大學
<a href="http://www.nctu.edu.tw/">國立交通大學 / National Chiao Tung University</a>         Oname: 國立交通大學
<a href="http://www.ntnu.edu.tw/">國立臺灣師範大學 / National Taiwan Normal University</a>   Oname: 國立臺灣師範大學
<a href="http://www.ncu.edu.tw/">國立中央大學 / National Central University</a>             Oname: 國立中央大學
<a href="http://www.nsysu.edu.tw/">國立中山大學 / National Sun Yat-sen University</a>       Oname: 國立中山大學
<a href="http://www.ncku.edu.tw/">國立成功大學 / National Cheng Kung University</a>         Oname: 國立成功大學
<a href="http://www.ccu.edu.tw/">國立中正大學 / National Chung Cheng University</a>         Oname: 國立中正大學
<a href="http://www.ym.edu.tw/">國立陽明大學 / National Yang Ming University</a>            Oname: 國立陽明大學
<a href="http://www.ndhu.edu.tw/">國立東華大學 / National Dong Hwa University</a>           Oname: 國立東華大學
<a href="http://www.ntou.edu.tw/">國立臺灣海洋大學 / National Taiwan Ocean University</a>    Oname: 國立臺灣海洋大學
<a href="http://www.ncnu.edu.tw/">國立暨南國際大學 / National Chi-Nan University</a>        Oname: 國立暨南國際大學
<a href="http://sun5.cpu.edu.tw/">中央警察大學 / Central Police University</a>              Oname: 警察大學
<a href="http://www.ntptc.edu.tw/">國立台北師範學院 / National Taipei Teachers College</a>   Oname: 國立台北師範學院
<a href="http://www.tmtc.edu.tw/">台北市立師範學院 / Taipei Municipal Teachers College</a>   Oname: 台北市立師範學院
<a href="http://www.nia.edu.tw/">國立藝術學院 / National Institute of the Arts</a>          Oname: 國立藝術學院
<a href="http://www.ntcn.edu.tw/">國立台北護理醫學院 / National Taipei College of Nursing</a> Oname: 國立台北護理醫學院
<a href="http://www.ntit.edu.tw/">國立台灣工業技術學院 / National Taiwan Institute of Technology</a>   Oname: 國立台灣工業技
    術學院
<A HREF="http://www.princeton.edu/index.html">普林斯頓大學</A>                        Oname: 普林斯頓大學
<a href="http://www.tccm.edu.tw/">慈濟醫學院 / Tzu Chi College of Medicine</a>              Oname: 慈濟醫學院
<a href="http://www.cyit.edu.tw/">朝陽技術學院 / Chaoyang Institute of Technology</a>       Oname: 朝陽技術學院
<a href="http://www.yzit.edu.tw/">元智工學院 / Yuan-Ze Institute of Technology</a>          Oname: 元智工學院
<a href="http://www.kpi.edu.tw/">高雄工學院 / Kaohsiung Polytechnic Institute</a>           Oname: 高雄工學院
<a href="http://www.chpi.edu.tw/">中華工學院 / Chung-Hua Polytechnic Institute</a>          Oname: 中華工學院
<a href="http://www.dyit.edu.tw/">大葉工學院 / Da-Yeh Institute of Technology</a>           Oname: 大葉工學院
<a href="http://www.ntcic.edu.tw/">國立臺北商業專科學校 / National Taipei College of Business</a>
                                                                                  Oname: 國立臺北商業專科學校
<a href="http://www.ntcic.edu.tw/">國立臺中商業專科學校 / National Taichung Institute of Commerce</a>
                                                                                  Oname: 國立臺中商業專科學校
<a href="http://www.nptic.edu.tw/">國立屏東商業專科學校 / National Pingtung Institute of Commerce</a>
                                                                                  Oname: 國立屏東商業專科學校
<a href="http://www.ncia.edu.tw/">國立嘉義農業專科學校 / National Chia-Yi Institute of Agriculture</a>
                                                                                  Oname: 國立嘉義農業專科學校
<a href="http://www.niiat.edu.tw/">國立宜蘭農工專科學校 / National Ilan Institute of Agriculture and Technology</a>
                                                                                  Oname: 宜蘭農工專科學校
<a href="http://www.nkit.edu.tw/">國立高雄工商專科學校 / National Kaohsiung Institute of Technology</a>
                                                                                  Oname: 國立高雄工商專科學校
<a href="http://www.ncit.edu.tw/">國立勤益工商專科學校 / National Chinyi Institute of Technology</a>
                                                                                  Oname: 國立勤益工商專科學校
<a href="http://www.lctc.edu.tw/">國立聯合工商專科學校 / National Lien-Ho College of Technology and Commerce</a>
                                                                                  Oname: 國立聯合工商專科學校
<a href="http://www.nypi.edu.tw/">國立雲林工業專科學校 / National Yunlin Polytechnic Institute</a>
                                                                                  Oname: 國立雲林工業專科學校
<a href="http://www.nkhc.edu.tw/">國立高雄餐旅管理專科學校 / National Kaohsiung Hospitality College</a>
                                                                              Oname: 國立高雄餐旅管理專科學校
<a href="http://ntcpe.ntcpe.edu.tw/">國立台灣體育專科學校 / National Taiwan College of Physical Education</a>
                                                                                  Oname: 國立台灣體育專科學校
<a href="http://www.ntcic.edu.tw/">臺南家政專科學校 / Tainan College of Home Economics</a>   Oname: 臺南家政專科學校
<a href="http://www.tccn.edu.tw/">佛教慈濟護理專科學校 / Buddhist Tz'u Chi Junior College of Nursing</a>
                                                                                  Oname: 慈濟護理專科學校
<a href="http://www.chs.edu.tw/">健行工商專校 / Chien Hsien Institute of Technology and Commerce</a>        Oname: 健行工商
<a href="http://www.vit.edu.tw/">萬能工商專科學校 / VanNung Institute of Technology</a>     Oname: 萬能工商專科學校
<a href="http://203.68.40.3/">南亞工商專科學校 / Nanya Junior College</a>                 Oname: 南亞工商專科學校
<a href="http://gopher.lhjc.edu.tw/">龍華工商專科學校 / Lunghwa Junior College of Technology and Commerce</a>
                                                                                  Oname: 龍華工商專科學校
```

<a href="http://www.mhit.edu.tw/">明新工商專校 / Ming Hsin Institute of Technology</a>	Oname: 明新工商
<a href="http://www.thctc.edu.tw/">大華工商專科學校 / Ta Hua College of Technology and Commerce</a>	Oname: 大華工商專科學校
<a href="http://www.chinmin.edu.tw/">親民工商專科學校 / Chin Min College of Technology and Commerce</a>	Oname: 親民工商專科學校
<a href="http://www.stjctc.edu.tw/">樹德工商專科學校 / Shu Teh Junior College of Technology</a>	Oname: 樹德工商專科學校
<a href="http://www.ccjc.edu.tw/">中州工商專校 / Chung Chou Junior College of Technology and Commerce</a>	Oname: 中州工商專校
<a href="http://203.64.144.1/">建國工商專科學校 / Chienkuo Junior College of Technology</a>	Oname: 建國工商專科學校
<a href="http://www.wfc.edu.tw/">吳鳳工商專科學校 / Wu-Feng Junior College of Technology and Commerce</a>	Oname: 吳鳳工商專科學校
<a href="http://www.ntc.edu.tw/">南台工商專科學校 / Nan Tai College of Technology and Commerce</a>	Oname: 南台工商專科學校

[Organization-Club (Oname)]

<a href="http://140.113.11.235/~gmusic/">台大佳韻音樂社</a>	Oname: 佳韻音樂社
<a href="http://cc.ntu.edu.tw/~b4101009/piano/">台大鋼琴社</a>	Oname: 鋼琴社
<a href="http://med.mc.ntu.edu.tw/~b0401087/chorus/">杏林合唱團</a>	Oname: 杏林合唱團
<a href="http://med.mc.ntu.edu.tw/~b3401006/sinlin/index.htm">杏林弦樂團</a>	Oname: 弦樂團
<a href="http://king.cc.ntu.edu.tw/~b1207031/">基克工作室</a>	Oname: 基克工作室

[Organization-Government (Oname)]

<a href="http://expo96.org.tw/">網路博覽會 中華民國館 / Pavilion of Taiwan, R.O.C.</a>	Oname: 中華民國館
<A HREF="http://expo96.org.tw/Welcome_c.html">中華民國館</A>	Oname: 中華民國館
<A HREF="http://www.motc.gov.tw/Welcome_c.html">交通館</A>	Oname: 交通館
<A HREF="http://www.nmns.edu.tw/">國立自然科學博物館 </A>	Oname: 國立自然科學博物館
<a href="http://www.nccu.edu.tw/zoo/htm/zoomain.htm">台北市立動物園 </a>	Oname: 台北市立動物園
<A HREF="http://192.192.14.202/welcome.htm">國立中正文化中心</A>	Oname: 國立中正文化中心
<A HREF="http://crab.ccl.itri.org.tw/cgi/m_normal">國家圖書館遠距圖書服務系統</A>	Oname: 國家圖書館

[Personal name (Pname)]

"http://dodger.ee.ntu.edu.tw/~lswang/">王立三的 HomePage / Li-San Wang's Homepage</a>	Pname: 王立三
"http://www.csie.ntu.edu.tw/~jcwang/index.cgi">王家俊 / John's House</a>	Pname: 王家俊
"http://med.mc.ntu.edu.tw/~shouzen/">生命的照顧 － 范守仁醫師 / Life Care - Fan's Home</a>	Pname: 范守仁
"http://king.cc.ntu.edu.tw/~d0701021/hgt/">何子之網頁</a>	Pname: 何子
"http://www.ee.ntu.edu.tw/~b82070/">杜立群</a>	Pname: 杜立群
"http://nlg3.csie.ntu.edu.tw/group/gwbian.html">邊國維的網頁</a>	Pname: 邊國維
"http://osil.csie.ntu.edu.tw/~chwu/">吳俊興</a>	Pname: 吳俊興
"http://king.cc.ntu.edu.tw/~b3401111/">吳振漢的窩 / Wilfred's HomePage</a>	Pname: 吳振漢
"http://king.cc.ntu.edu.tw/~b3502118/">林育德（AirL)的遊園地</a>	Pname: 林育德
"http://king.cc.ntu.edu.tw/~b2504049/">林欣蔚 / CELHW</a>	Pname: 林欣蔚
"http://ipmc.ee.ntu.edu.tw/~sclin/">林信成的 W3 小棧</a>	Pname: 林信成
"http://king.cc.ntu.edu.tw/~b2501109/welcome.htm">依客那米克斯傳說—勇者耀耀之章</a>	Pname: 那米克斯
"http://140.112.19.6:8000/">阿哲的夢幻天地</a>	Pname: 阿哲
"http://med.mc.ntu.edu.tw/~green/">林錦鴻 - 電腦玩家，網路流民，婦產科醫師</a>	Pname: 林錦鴻
"http://king.cc.ntu.edu.tw/~b2501127/">唐唐的世界</a>	Pname: 唐唐
"http://king.cc.ntu.edu.tw/~b2603230/">張正宜-不來不可的好地方 / TOM's Home</a>	Pname: 張正宜
"http://sun.gcc.ntu.edu.tw/Huang/">黃兆談</a>	Pname: 黃兆談
"http://king.cc.ntu.edu.tw/~r5241206/">魚兒的小鎭ー林康捷的 Homepage</a>	Pname: 林康捷
"http://king.cc.ntu.edu.tw/~b3503015/">陳紀光 / HomePage of Chen Chi-kuang</a>	Pname: 陳紀光
"http://cml19.csie.ntu.edu.tw/~robin/">陳炳宇 / Robin's Workgroup</a>	Pname: 陳炳宇
"http://med.mc.ntu.edu.tw/~b9401011/">郭昇彥的烘焙機</a>	Pname: 郭昇彥

Appendix C. Some Mapping Results in the Content Part

In the following, Oname and Pname denote the extracted organization names and personal names, respectively. The number indicates the token no. of the information in the web pages.

[Some Extracted Data in Content Sets before Mapping]

Oname: 資訊新館 63
E-Mail: root@csman.csie.ntu.edu.tw 69

Oname: 土木館 81
E-Mail: root@ce.ntu.edu.tw 82

Pname: 蔡博文 108
Oname: 地理系館 109
E-Mail: tsaibw@ccms.ntu.edu.tw 112

Pname: 丘台生 122
Oname: 漁科館 123
E-Mail: tschiu@ccms.ntu.edu.tw 124

Pname: 陳膺州 146
E-Mail: ingchen@chem60.ch.ntu.edu.tw 152

Pname: 張震東 155
E-Mail: gdchang@ccms.ntu.edu.tw 160

Pname: 黃靜美 171
E-Mail: mei@ccms.ntu.edu.tw 175

Pname: 林翰彥 178
Oname: 森林館 179
E-Mail: wenliang@ccms.ntu.edu.tw 180

Pname: 蘇明道 184
Oname: 農工館 185
E-Mail: sumd@ccms.ntu.edu.tw 186

Pname: 王友俊 382
E-Mail: wangecaa@ccms.ntu.edu.tw 387

Pname: 周伯戢 389
E-Mail: pkchou@ccms.ntu.edu.tw 391

Pname: 游張松 250
E-Mail: yucs@ccms.ntu.edu.tw 254

Pname: 曾珀雯 270
Pname: 徐信權 272
E-Mail: popo@ccms.ntu.edu.tw 276
E-Mail: kevins@ccms.ntu.edu.tw 277

[Some Mapping Results in Content Sets]

E-Mail: root@csman.csie.ntu.edu.tw	Oname: 資訊新館
E-Mail: focus@www.ntu.edu.tw	Oname: 焦點新聞
E-Mail: news@www.ntu.edu.tw	Oname: 網路新聞
E-Mail: campus@www.ntu.edu.tw	Oname: 校園新聞
E-Mail: tsaibw@ccms.ntu.edu.tw	Pname: 蔡博文
E-Mail: tschiu@ccms.ntu.edu.tw	Pname: 丘台生
E-Mail: ingchen@chem60.ch.ntu.edu.tw	Pname: 陳膺州
E-Mail: yucs@ccms.ntu.edu.tw	Pname: 游張松
E-Mail: hlee@cc.ntu.edu.tw	Pname: 李賢輝
E-Mail: popo@ccms.ntu.edu.tw	Pname: 曾珀雯
E-Mail: kevins@ccms.ntu.edu.tw	Pname: 徐信權
http: http://www.ntu.edu.tw/forest/R17.html	Oname: 國立臺灣大學森林學系暨研究所

Computational Linguistics and Chinese Language Processing
vol 3. no.1 February 1998, pp. 101-116
© Computational Linguistics Society of R.O.C.

Human Judgment as a Basis for Evaluation of Discourse-Connective-Based Full-Text Abstraction in Chinese

Benjamin K T'sou[*], Hing-Lung Lin[*], Tom B Y Lai[*], Samuel W K Chan[*]

Abstract

In Chinese text, discourse connectives constitute a major linguistic device available for a writer to explicitly indicate the structure of a discourse. This set of discourse connectives, consisting of a few hundred entries in modern Chinese, is relatively stable and domain independent. In a recently published paper [T'sou 1996], a computational procedure was introduced to generate the abstract of an input text using mainly the discourse connectives appearing in the text. This paper attempts to demonstrate the validity of this approach to full-text abstraction by means of an evaluation method, which compares human efforts in text abstraction with the performance of an experimental system called ACFAS. Specifically, our concern is about the relationship between the perceived importance of each individual sentence as judged by human beings and the sentences containing discourse connectives within an argumentative discourse.

Keywords: text abstraction, discourse connectives, performance evaluation, experiment design, correlation analysis

1. Introduction

As a result of increasingly convergent interests and cross-fertilization in linguistics and computer science, research into discourse in natural language processing (NLP) has made much progress in the last decade. Discourse as understood by linguists refers to any form of language-based purposeful communication involving multiple sentences or utterances. The most important forms of discourse of interest to NLP are text and dialogue. While discourse, either textual or spoken, normally appears as a linear sequence of sentences, it has long been recognized by linguists that these sentences tend to cluster together into units, called discourse segments, that are related in some way and form a hierarchical

[*] Language Information Sciences Research Center, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong. E-mail: rlbtsou@cpccux0.cityu.edu.hk

structure.

In NLP, discourse analysis must go beyond sentence-based syntactic and semantic analysis. Its functions are to divide a text into discourse segments and to recognize and re-construct the discourse structure of the text as intended by its author [Allen 1995]. Results of discourse analysis can be used to resolve many important NLP problems, such as anaphoric reference [Hirst 1981], tense and aspect analysis [Hwang 1992], intention recognition [Grosz 1986, Litman 1990], text generation [McKeown 1985, Lin 1991] etc.

Discourse analysis is also applicable to text abstraction, as demonstrated in Project ACFAS (Automated Chinese Full-text Abstraction System), which aims to automatically produce abstracts from Chinese newspaper editorials published in Hong Kong [T'sou 1992, T'sou 1996] using a new approach based on analyzing the rhetorical structure of argumentative discourse. This process, called Rhetorical Structure Analysis (RSA) [T'sou 1996], is based on the Rhetorical Structure Theory developed by Mann and Thompson for describing the discourse structure of English text [Mann 1986]. A similar approach has been applied to Japanese [Ono 1994].

As a brief review of the RSA process, please note that in modern Chinese text, discourse connectives constitute a major linguistic device available to a writer to explicitly indicate the structure of a discourse. Examples of Chinese discourse connectives include 因此 ("therefore"), 因爲 ("because"), 如果 ("if")... 就 ("then"), 假如 ("assuming")... 那末 ("then"), 雖然 ("although") …但是 ("but") etc. This set of discourse connectives, consisting of a few hundred entries, is relatively stable in modern Chinese and is independent of the domain of discourse. Initial corpus analysis [Ho 1993] has indicated that about 30% of the clauses in a typical Chinese editorial published in a Hong Kong newspaper contain explicit discourse connectives, which are used to express the temporal, causal or rhetorical relationships amongst constituent propositions or clauses.

The principle of the RSA process is predicated on the fact that the appearance of discourse connectives in a Chinese argumentative text, such as a newspaper editorial, constitute a key to the basic understanding of the inherent logical structure underlying its argumentative discourse and, thus, provide a potentially useful approach to scaleable and domain-independent full-text abstraction. Generally speaking, the RSA process makes use of those discourse connectives appearing in a Chinese text to (1) extract every rhetorically connected discourse segment of text and (2) recognize and construct the rhetorical structure of each discourse segment. Using the resultant disconnected rhetorical structures, an appropriate abstract can be generated by means of systematic rhetorical structure reduction to produce abstracts with differential coverage of the details of the underlying argumentation (for details of the algorithm, please refer to [T'sou 1996]).

Because the flow of argumentation is not exclusively demarcated by discourse connectives, the validity and robustness of this approach require empirical comparison with human efforts in abstraction, which can contribute to the design of a general evaluation method for automatic abstraction in Chinese. Such a comparison would entail human subjects performing abstraction on the same editorials used in ACFAS and comparing their results (see also [Watanabe 1996]). Two major questions require answers obtained from carefully designed experiments: (1) Is there relative consistency in human abstraction? (2) Is the existence of discourse connectives a relevant factor in determining the relative importance of constituent discourse segments?

As a preview, Section 2 describes how the experiments were conducted with emphasis on the rationale behind the design of these experiments. In Section 3, we delineate our method of analysis and present the formal definitions of the evaluation metrics. In Section 4, we show that, based on our experimental results, abstracts produced by different groups of human subjects with similar educational background are relatively consistent when they are examined as a group. Section 5 reports the results of performance evaluation of ACFAS using the metrics of recall and precision. In our conclusion, we stress the importance of systematic and quantitative evaluation of various factors that can contribute to the design of automated full-text abstraction systems.

2. Design of the Experiment

A set of 10 Chinese editorials was taken from two well-known newspapers published in Hong Kong and denoted as {E1, E2, ⋯ ., E10}. These editorials were concerned with controversial events which occurred in Hong Kong. They included a decision to build a nuclear power plant near Hong Kong, the relationship between debt and corruption in the police force, the unemployment rate of young people, the law and the attitude of the population towards anti-discrimination etc. These editorials are arche-typical examples of argumentative discourse.

The subjects of the experiment included three groups of 25 students each from three prestigious universities in northern China. Two groups were from Chinese departments and one was from a computer science department; all the students were either final year undergraduates or first year graduate students. They participated in the experiment separately in time and location; as far as we can ascertain, these were independent experiments. The subjects were generally brought up in primarily monolingual settings and could understand the issues discussed in the selected editorials but without intimate knowledge or prejudice with regard to the related background. It was our conscious decision to use Hong Kong newspaper editorials with Mainland Chinese subjects of above-average linguistic competence and intellectual capacity for the sake of performance comparison.

Computer print-outs instead of the original texts were given to the subjects of this experiment to avoid any confusion or hints preserved in the format of the original texts. The experiments were conducted under a controlled environment in an invigilated classroom setting.

The subjects were given the 10 selected editorials in one batch. They were asked to determine which clauses or sentences in each given editorial contained the most essential information provided by the author. The subjects were required to work on the editorials sequentially and within a prescribed amount of time. Each subject was asked to (1) underline in red about 10% of text which, according to his/her own judgment, contained the most important information (called key propositions below) in the editorial, and (2) underline in blue about 15% more of the next most important parts (called important propositions below) of the editorial. The subjects were specifically advised to cover as widely as possible (subject to the above constraints, of course) all aspects of the content that the author might have intended to convey.

After the experiments were conducted, the importance of each proposition was evaluated on the basis of how the text was marked by the subjects of each experiment according to the method discussed in the following section.

3. Method of Analysis and Evaluation Metrics

Data analysis of the experimental results as well as performance evaluation of ACFAS were carried out as follows: (1) Target abstracts were generated per editorial per student group according to how the editorial text was marked by the human subjects. (2) Target abstracts for the same editorial were analyzed for similarity and consistency among the three groups. (3) Abstracts generated by ACFAS were compared with the corresponding abstracts generated by the human subjects according to two performance metrics, recall and precision, as defined in Section 3.2.

3.1 Generation of the target abstract

The objective of this step was to select part of a given source text to form a target abstract. The selection criterion was based on how the text was marked by the human subjects of the experiment.

(i) Let WK be the weighting factor assigned to a *key proposition*, and let
WI be the weighting factor assigned to an *important proposition*,
where $0 < WK, WI \leq 1$.

We can compute the weighted average of the jth proposition, denoted as PERC-IMPj (for *Perceived Importance*), according to the following formula:

$$\text{PERC-IMP}_j = \frac{1}{n}\left\{ (\sum_{i=1}^{n} \text{KEY}_{ij}) * \text{WK} + (\sum_{i=1}^{n} \text{IMP}_{ij}) * \text{WI} \right\},$$

where n is the number of subjects,

$$\text{KEY}_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ proposition is marked by the } i^{th} \text{ subject as a key proposition,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{and } \text{IMP}_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ proposition is marked by the } j^{th} \text{ subject as an important proposition,} \\ 0 & \text{otherwise} \end{cases}$$

(ii) For a given source text, we can sort all the propositions of the text according to their perceived importance.

Let α ($0 < \alpha \leq 1$) be the threshold value used to separate those propositions that should be included in the *target abstract* (for $\text{PERC-IMP}_j \geq \alpha$) and those that should be excluded (for $\text{PERC-IMP}_j < \alpha$). Note that α is introduced to account for the fact that, when we talk about abstraction of a source text, there is a whole spectrum of possible abstracts with different sizes, each of which corresponds to a different value of α.

For a given α, we can define the *abstract ratio*, β, of the target abstract as

$$\beta(\alpha) = \frac{\text{size of target abstract}(\alpha)}{\text{size of source text}}$$

3.2 Performance metrics for a text abstraction system

ACFAS is an experimental text abstraction system that is capable of generating multiple abstracts with differential coverage of a source text. In this study, we consider only the abstract generated by the top-level output of ACFAS [T'sou 1996]. We define the abstract to source ratio of the top-level output of ACFAS as

$$\text{ACFAS-RATIO} = \frac{\text{size of top} - \text{level abstract of ACFAS}}{\text{size of source text}}$$

The following two performance measures for ACFAS are defined:

$$\text{RECALL}(\beta) = \frac{\text{\# of target propositions generated by ACFAS}}{\text{size of target abstract}}$$

$$\text{PRECISION}(\beta) = \frac{\text{\# of target propositions generated by by ACFAS}}{\text{size of abstract generated by ACFAS}}$$

Note that in the above definitions, target propositions are propositions that are included in the target abstract as defined in Subsection 3.1. Since every target abstract is a function of some threshold value α such that the perceived importance of each target proposition in it is greater than or equal to α, to conduct an evaluation, we are generally interested in controlling the size of the target abstract and not the specific value of α. Therefore, we explicitly indicate that both RECALL and PRECISION are functions of the abstract ratio β of the target abstract that we choose for conducting an evaluation.

4. Similarity Analysis of Human-Generated Abstracts

In this section, results of the experiment described in Section 2 above are analyzed within the framework set out in Section 3 to examine consistency in abstracts generated by different groups of human subjects.

Text abstraction is the process of condensing salient information from a source text. It involves sophisticated and intelligent manipulation of given and assumed world knowledge as well as knowledge of natural language. It is well known that abstracts produced by different human individuals from the same source text can vary depending on, for example, the background and education level of the individuals involved. Furthermore, even for the same individual, different abstracts can be generated at different times [Luhn 1958]. While this is true with respect to the behavior of individual human beings, when they are examined as a group, our results below show that abstracts produced by different groups of human subjects with similar educational background in a given society are in fact relatively consistent. This result shows that there is an aspect of consistency in human summarization of text, which can provide a basis for evaluation of automated text abstraction systems. Further research in psychological studies is required to explore the cognitive basis for this.

Fig. 1 shows the average Perceived Importance scores for the 65 propositions in one of the test editorials with respect to each group of subjects. In the appendix, we show the same editorial divided into individual propositions for the reader's reference. The two weighting factors are set to be WI=0.8 and WK=1. These two values are chosen to reflect the fact that key propositions and important propositions constitute the top 10% and the next 15%, respectively, of the source text according to the instruction given to the subjects of the experiment.

Inspection of the three plots shown in Fig. 1 reveals that while there is considerable variation in the (three) absolute scores of each of the individual propositions, the overall shapes of the three plots are obviously similar.

The similarity of the plots was statistically assessed by considering each of the propositions as an observation point. For the sake of convenience, the scores given by the 25 subjects in a group were averaged, so that there were 3 scores for each of the observation points. Pearson coefficients of correlation (pair-wise) of the (averaged) scores of the three groups calculated from data for 379 propositions (or observation points) in 5 common test editorials are given in Table 1.

As shown in Table 1, the correlation coefficients are positive and close to 1. They clearly establish strong consistency amongst the three groups of human subjects with respect to their perception of the relative importance of individual propositions in the editorials. Besides confirming that human subjects do indeed generate abstracts in a consistent manner, the above analysis can also be seen as empirical evidence of the validity of the Perceived Importance score suggested in Section 3.

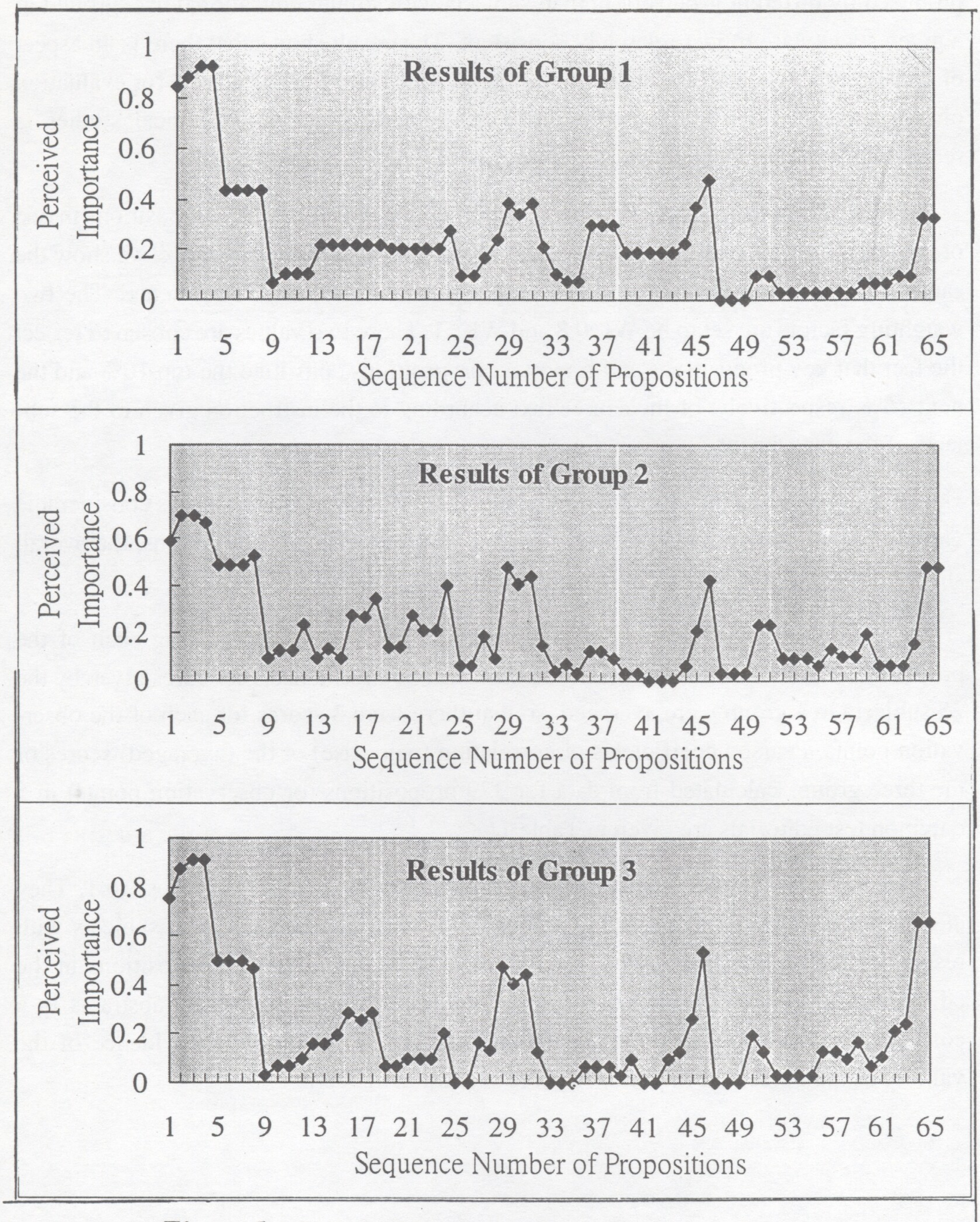


Figure 1 *Perceived Importance of an Editorial for Three Groups of Subjects*

	Group 1	Group 2	Group 3
Group 1	1		
Group 2	0.886077	1	
Group 3	0.914838	0.945098	1

Table 1. *Coefficients of Correlation of the Perceived Importance Scores of 5 Editorials for Three Groups of Subjects*

5. Performance Evaluation of ACFAS: An Empirical Study

In the previous section, we demonstrated that abstracts generated by different groups of human subjects exhibit a high degree of similarity. Therefore, it seems appropriate to evaluate the performance of a text abstraction system by comparing its output with target abstracts produced by human subjects based on the metric of Perceived Importance. In this section, we report the results of an empirical study on the performance of ACFAS based on the performance measures RECALL and PRECISION defined in Section 3. This evaluation was conducted by comparing abstracts generated by ACFAS with target abstracts produced by a group of 25 computer science students.

5.1 Statistics on the target abstracts of 10 source texts

The average target abstract ratios of 10 editorials, given as a function of the Perceived Importance threshold, are shown in Figure 2. The two weighting factors were set to be WI=0.8 and WK=1 as discussed above. On average, only 12.5% of the contents of any source text received a Perceived Importance of 0.5 or above. This indicates that, within any text, there exists a small, identifiable group of propositions which contains the most important information relevant to the text. This small group of propositions will form the basis of any abstract produced by human subjects.

On the other hand, it may be noted that about 40% of the content of any source text received a Perceived Importance of less than 0.1. This very likely indicates a high degree of redundancy in human compositions of this genre.

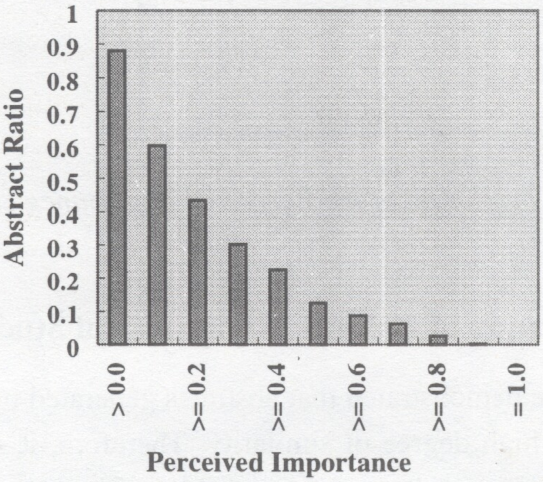


***Figure 2** Abstract Ratio as Function of Perceived Importance*

5.2 Statistics on top-level abstracts generated by ACFAS

On average, the size of a top-level abstract generated by ACFAS was 27.4% of the source text. This is significantly higher than the target abstract ratio of 12.5% (for $\alpha \geq 0.5$) produced by human subjects. This result may be caused by a lack of explicit discourse connectives needed to determine the relationships between different (yet related) discourse segments. An in depth study on more general types of discourse connectives, including explicit and implicit ones, should improve the present situation.

5.3 Performance evaluation of ACFAS

The average RECALL and PRECISION of the 10 abstracts generated by ACFAS, according to how well they correspond with the target abstracts produced by human subjects, are shown in Fig. 3 and 4.

As shown in Fig. 3, when the abstract ratio (i.e., the human-generated abstract size as a percentage of the source text) equals to 100%, the average RECALL is 27.4%, which is also the size of the top-level abstract generated by ACFAS. As the value of the abstract ratio is reduced, the average RECALL increases modestly until it reaches a maximum value of 36.5% for an abstract ratio of 30%. This improvement of about 10% for average RECALL is an indication of an inherent relationship between the mechanism of ACFAS and the process of human text abstraction.

Note that when the abstract ratio of 30% is further reduced, the average RECALL decreases rapidly. As our abstract ratio is computed by sorting all the propositions in the text according to their perceived importance, a small abstract ratio corresponds to the set of propositions that have received high average scores of perceived importance. This result indicates that ACFAS is unable to retrieve some of the most important propositions from the text. After examining the content of the source texts, we find that there is a high probability of finding important propositions at the beginning and the end of these texts (which seems to reflect a typical pattern in argumentative discourse, i.e., a problem statement at the beginning and conclusion at the end of a text), but relatively few discourse connectives are found in this area. The present strategy of ACFAS is to ignore sentences without explicit discourse connectives between them; therefore, those target propositions located at the beginning and the end of the text will not be included in the ACFAS-generated abstract.

Fig. 4 contrasts the values of RECALL and PRECISION, both as functions of the abstract ratio. We observe that at the maximum RECALL rate of 36.5%, the average PRECISION is 39.4%. In other words, about 60% of the target propositions are not extracted by ACFAS, and most of them are propositions located at the beginning and end of the source texts.

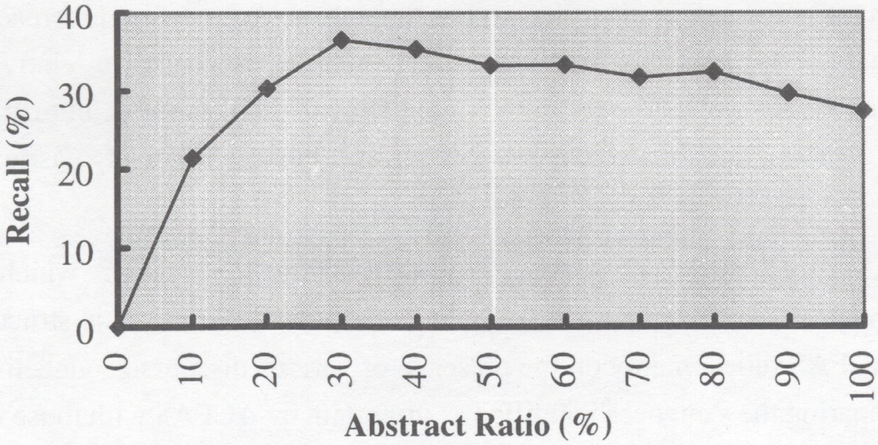


Figure 3 *Recall as a Function of Abstract Ratio*

The conclusion we can draw from this result is that a system like ACFAS, which uses only the existence of explicit discourse connectives to determine the relative importance of the propositions in an argumentative discourse, performs well on the part of the text that deals with the argumentative flow and presentation of evidence but performs poorly where the problem statement is delineated and the conclusion or summary is presented. Other factors and cues must be used to account for this deficiency.

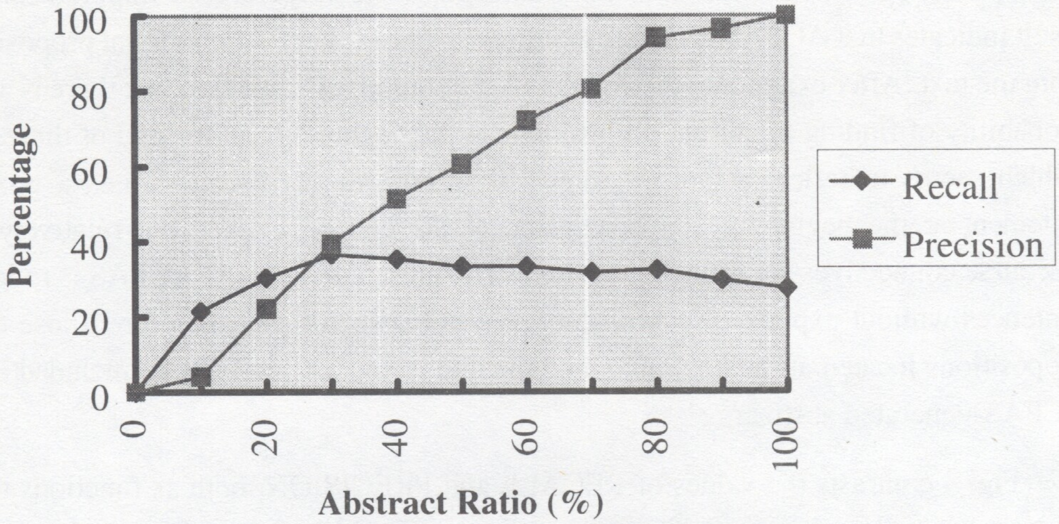


Figure 4 *Recall vs. Precision as Functions of Abstract Ratio*

6. Conclusions

Text abstraction entails the process of determining which sentences in a text contain the most important information that the author intends to convey to his readers. Our empirical study shows that this set of essential sentences consists of a relatively small fraction of the original text. Based on their comprehension of the text, human subjects, behaving as a group, are able to pinpoint this set of sentences relatively easily and consistently.

ACFAS is an automated Chinese full-text abstraction system, which extracts essential sentences from a given text following analysis of its discourse structure. This process in ACFAS relies mainly on the presence of various discourse connectives in the text. By comparing the sentences identified as important by ACFAS with those identified by human subjects, *who presumably use additional cues*, our study shows that there is a non-random correspondence between these two sets of sentences. Since ACFAS, in its current design, does not employ deep semantic processing to understand the meaning of each sentence in a text, we can conclude the following: Which information in a text is perceived by its readers as important depends not only on its semantic content, but also on how it is presented in the text, i.e., its discourse structure.

As a final remark, text abstraction represents a unique human faculty, which involves intelligent manipulation of given and assumed knowledge and natural language. Therefore, it is our belief that no single factor can guarantee its successful execution. Relevant factors or cues that have been used in the design of automated text abstraction systems include keywords, word frequency counts, discourse connectives, rhetorical relations, tense, and distance from the beginning and the end of a text, just to name a few. However, there has been general neglect·of systematic and quantitative evaluation of the relative contribution of each individual factor to the whole process of text abstraction. The present paper, by concentrating on the factor of explicit discourse connectives within a text, is a step toward improving this situation.

References

Allen, J., *Natural Language Understanding*, 2nd Edition, Reading, Benjamin/Cummings, Redwood City, CA, 1995.

Grosz, B.J. and C. Sidner, "Attention, Intention, and the Structure of Discourse," *Computational Linguistics* 12:3, 1986, pp.175-204.

Hirst, G., "Discourse Oriented Anaphoral Resolution in Natural Language Understanding: A Review," *Computational Linguistics* 7:2, 1981, pp. 85-98.

Ho, H.C., B.K. T'sou, Y.W. Chan, B.Y. Lai and S.C. Lun, "Using Syntactic Markers and Semantic Frame Knowledge Representation in Automated Chinese Text Abstraction," in *Proc. 1st Pacific Asia Conf. On Formal and Computational Linguistics*, Taipei, 1993, pp. 122-131.

Hwang, C.H. and L.K. Schubert, "Tense Trees as the 'Fine Structure' of Discourse," in *Proc. 30th Annual Meeting, Assoc. for Computational Linguistics*, 1992, pp. 232-240.

Lin, H.L., B.K. T'sou, H.C. Ho, T. Lai, C. Lun, C.K. Choi and C.Y. Kit, "Automatic Chinese Text Generation Based on Inference Trees," in *Proc. ROCLING Computational Linguistic Conf. IV*, Taipei, 1991, pp. 215-236.

Litman, D.J. and J. Allen, "Discourse Processing and Commonsense Plans," in Cohen et.al.(ed.), *Intentions in Communications*, 1990, pp. 365-388.

Luhn, H.P., "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, 2:2, 1958, pp. 159-165.

Mann, W.C. and S.A. Thompson, "Rhetorical Structure Theory: Description and Construction of Text Structures," in Kempen(ed.) *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, 1986, pp. 279-300.

McKeown, K.R., "Discourse Strategies for Generating Natural-Language Text," *Artificial Intelligence* 27:1, 1985, pp. 1-41.

Ono, K., K. Sumita and S. Miike, "Abstract Generation based on Rhetorical Structure Extraction," *Proc. Coling'94*, 1994, pp. 344-348.

T'sou, B.K., H.L. Lin, H.C. Ho and T. Lai, "From Argumentative Discourse to Inference Trees:

Using Syntactic Markers as Cues in Chinese Text Abstraction," in *Proc. 3rd International Conf. On Chinese Information Processing,* Beijing, China, 1992, pp. 76-93. Also appeared in C.R. Huang, K.J. Chen & B.K. T'sou (ed.) *Readings in Chinese Natural Language Processing*, Monograph Series No. 9, *Journal of Chinese Linguistics*, 1996, pp. 199-222.

T'sou, B.K., H.L. Lin, H.C. Ho, T. Lai and Terence Chan, "Automated Chinese Full-text Abstraction Based on Rhetorical Structure Analysis," *Computer Processing of Oriental Languages* 10:2, 1996, pp. 225-238.

Watanabe, H., "A Method for Abstracting Newspaper Articles by Using Surface Clues," *Proc. Coling'96*, 1996, pp. 974-979.

Appendix

The following editorial was originally published in the Sing Tao Morning Post on 20th December, 1995 entitled " 物業市道可望穩健發展 ". The proposition numbers are assigned by the authors of this paper for convenience of analysis. The discourse connectives recognized by ACFAS are underlined.

1 港府九五年最後一次土地拍賣的結果，
2 既反映出發展商對九七年後香港經濟有信心，
3 亦顯示商業樓市仍會繼續調整，
4 而豪宅、中下價樓宇則會個別發展。
5 港府的壓抑樓市措施、
6 本港經濟放緩，
7 以及中國實施宏觀調控這些因素，
8 使近兩年的賣地收入較預期爲低。
9 雖則港府高官表示樓價快將見底，
10 但地產市道還未復甦，
11 買賣仍然偏淡，
12 展望來年，本港經濟景氣低迷仍難望有重大改善。
13 故此，私人發展商將推出的樓盤雖是過往十年來最少的一年，
14 但政府則會有大量居屋供應，
15 這樣必然拖慢物業市道的復甦，
16 樓市可能較今年活躍，
17 卻不能與過往旺盛期間同日而語，
18 最多只能如發展商所預期的「穩健發展」。
19 因爲樓價雖回落到九三年的水平，
20 畢竟與一般市民的收入仍有一段距離，
21 而中國仍無意放寬宏觀調控，
22 以及利息縱使回落，
23 幅度亦不太大，

24 都使樓市不可能在短期內再創高峰。
25 昨天拍賣的半山區一幅興建豪宅的土地，競投熱烈，
26 與過去幾次的情況相若，
27 這是因市場對豪宅的需求仍殷，
28 地產發展商相信香港在九七年後仍能保持這一地區的經濟中心地位所致。
29 近期的樓市轉趨活躍與過往情況有別，
30 不是由中小型住宅帶動的，
31 而是由豪宅引起的，
32 其原因是豪宅供應有限，
33 最近十年來平均增加二千個單位，
34 明年的供應量，照政府估計，只及今年的一半，
35 約爲一千二百個。
36 因此，只要香港能吸引外資來投資，
37 則這類物業就不怕沒有租客，
38 租金亦會易升難跌，
39 故在樓價進入鞏固期後，
40 不少投資者買下這類物業作長線投資，
41 投得司徒拔道地段的淘大置業，
42 就計畫在樓宇建成後將之出租，
43 這與一般投資者大同小異，
44 都是對香港前途有信心而看好豪宅後市的一種表現。
45 但是，銅鑼灣渣甸坊一幅非工業用地的拍賣經過及成交價，
46 再度證實商業樓宇的後市仍不被看好。
47 固然，這塊地皮受到周圍環境所影響，
48 面積又不大，
49 落成後難與同區其他商廈匹敵，
50 但是更根本的原因，應是商廈的市場供應遠超於市場需求，
51 目前空置率偏高。
52 由於本港經濟放緩，
53 許多商戶爲減少營運成本，
54 將辦公地點搬往租金較廉宜的地區，
55 甚至是搬往他地。
56 另一方面，年前大陸投資者大舉入市炒賣甲級寫字樓，
57 使樓價高不可攀，
58 以致缺少承接力。
59 這兩個因素，使商業樓宇的價格及租金全面向下，
60 迄今已跌了四成，
61 雖然再大幅下挫的機會不大，
62 但是　由於經濟尙難復甦，
63 商廈市道在短期內亦難有轉機。
64 不過，這次賣地的氣氛與成績，多多少少顯示投資者對本港物業市道已較年初更有

信心，
預期來年可有穩步發展。