University of Alberta

An Unsupervised Learning Method for Cluster Labeling

by

Theresa Margaret Jickels

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science.**

Department of Computing Science

Edmonton, Alberta
Spring 2005

To my family

Abstract

Gathering semantic knowledge is important in many areas of Natural Language Processing. Semantic knowledge is particulary needed in Ontology Construction and Question Answering. In this thesis, a new unsupervised learning algorithm is developed to label groups of nouns with hypernym labels. The algorithm is feature-based, using syntactic features to train scores for each feature and then select possible labels. Selecting hypernym labels with this algorithm does not require annotated data. For evaluation, comparison is made between a baseline, the basic algorithm, variations, and WordNet. During evaluation, the methods are judged against human labeling decisions.

Acknowledgements

Thanks to Dr. Dekang Lin for the data and use of the LaTaT system.

Also a huge thanks to my supervisor, Dr. Greg Kondrak, for his insight, guidance, and patience.

I would also like to thank the participants of my study for contributing their valuable time to labeling so many clusters for me.

As well, thank you to my family for their support.

Contents

1 Introduction 1

2 Background 4

2.1 Definition of Common Terms 4

2.2 Tools 6

2.2.1 Clustering by Committee 6

2.2.2 WordNet and QueryData 6

3 Related Work 8

3.1 Dictionary Based Methods 8

3.2 Text Mining Methods 9

3.3 Named Entity Methods 9

3.4 Cluster-based Methods 10

4 The Technique 12

4.1 Motivation 13

4.2 Terminology 13

4.3 Data Preparation 15

4.4 System Overview 16

4.5 Training the Model 17

4.6 Variations on the Basic Method 20

4.6.1 Frequency Filtering 22

4.6.2 Baseline Method 22

4.6.3 Heuristic methods on the feature data 23

4.6.4 Named Entities 24

4.6.5 Making Use of WordNet 25
4.6.6 Combining Feature Score and WordNet 25

5 Experimental Setup and Results **26**
5.1 Experimental Setup 26
5.2 Human Performance Experiment 28
5.3 Development Set Results 30
5.3.1 Baseline Experiment 30
5.3.2 Heuristic Experiment 32
5.3.3 Basic Method Experiment 32
5.3.4 WordNet experiment 35
5.3.5 Frequency Filtering Experiment 35
5.3.6 Intersection Experiment 36
5.4 Test Set Results 37
5.4.1 Baseline Experiment 37
5.4.2 Heuristic Experiment 37
5.4.3 Basic Method Experiment 38
5.4.4 WordNet Experiment 40
5.4.5 Frequency Filtering Experiment 40
5.4.6 Intersection Experiment 42
5.5 Discussion of Experimental Results 42

6 Conclusions and Future Work **44**

Bibliography **46**

A Data example **48**

B Feature List **50**

C LaTaT Script **53**

D The Test Clusters **54**

E Aggregate Human Results **70**

List of Tables

4.1 The FS'(f) scores after the loop 20
4.2 The LS(a) scores after convergence 20

List of Figures

1.1 A small example of a possible network 2

4.1 Sample dependency data for a noun group 12
4.2 Examples of Terms 14
4.3 The Training Algorithm 18
4.4 The Feature Score Training Example: Initialization 19
4.5 The Feature Score Training Example: Final Values 21
4.6 The Feature Sets 23
4.7 Answer to Feature Distribution 24

5.1 Results of Experiments on the Development Set 31
5.2 The Baseline and Basic Training Method Results on Named and Regular Entities on the Development Set 33
5.3 Results on the Test Set 39
5.4 Results of Regular and Named parts of the Test Set 41

Chapter 1

Introduction

In many areas of computing science, and also in linguistics, references to ontologies, semantic networks and lexical databases can be found. These resources all use similar knowledge in their construction. The problem of producing them is a difficult one. It has been an object of study for some time, and in general, the options are to build such resources by hand (a time consuming project at best) or to automatically gain and organize the semantic information using a computer. Hand-built knowledge bases take much effort and time to produce, while efforts to automatically extract the information have met with varying degrees of success.

Artificial Intelligence topics and their potential solutions often require some form of knowledge representation. One type of representation is the ontology. Since the terms semantic network, ontology, and lexical database refer to similar representations, for the purposes of this work they will be referred to equivalently. An ontology is an organization of concepts and their properties. The structure of ontology enables some deductions about the concepts represented in it. For example, figure 1.1 shows a part of a possible ontology network. In this case, it can be deduced that dogs have tails, and that golden retrievers are animals, for example. In order to build such knowledge bases, the semantic relations (such as golden retriever being a type of dog) need to be available.

Since semantic relations are what ontologies and lexical databases are built on, the generation of this information can be used to either expand or build

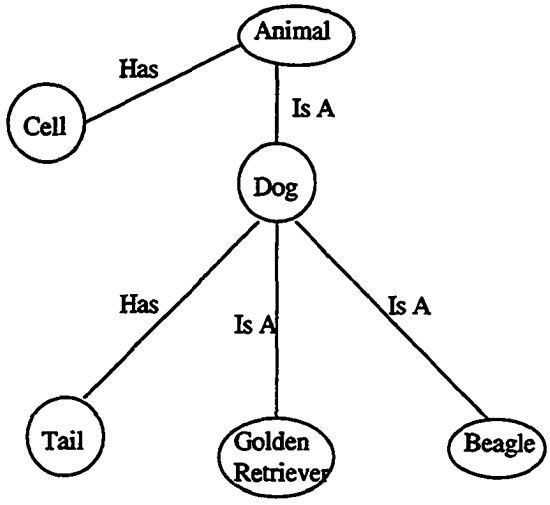

Figure 1.1: A small example of a possible network

new ontologies, semantic networks and lexical databases. The automatic and unsupervised aspect of the algorithm presented in this thesis means that the cost of producing these ontologies will be less than that of a hand-built or supervised technique. The reason for the savings is that the hand-built system would have to be constructed by humans, and supervised techniques are algorithms that learn to solve a problem by observing the answers, so they would require a set of data with the answers (which would have to be made by humans). An example of a hand-built ontology is WordNet, created by Miller et al. [8], which is in frequent use.

This work addresses a specific subset of the problem of automatic gathering of semantic relations by computer. The goal of this work is to develop an unsupervised algorithm for the purpose of categorizing groups of terms. Since the algorithm is unsupervised, it can produce categories quickly, and without the prohibitive cost of human involvement in annotating training data. Here the set of terms (*table, chair, desk, sofa, dresser, bookcase*) is labeled with the term *furniture*, since all the items in the set are types of furniture. A second example is the group of titles (*Deep Impact, Lethal Weapon 4, Armageddon, Godzilla, Titanic, Truman Show, Mulan*). For this second set of terms, possible labels include *movie*, or *film*. The relationship between these terms and their category is known as a hypernym relationship (i.e. *furniture*

is a hypernym of *table*).

There are many possible applications of the proposed algorithm in applications of Computing Science such as Question Answering, and Named Entity Classification. All of these contribute to the field of Natural Language Processing. The algorithm presented here gives a new, unsupervised method to produce useful semantic relationships, contributing towards these goals.

One area in which these semantic relations is useful is that of Question Answering. Question Answering refers to the problem of having a computer find the answer to questions from a body of text. Many common question types can utilize the semantic relations produced by such categorizing mentioned above (in fact WordNet [8] is used quite frequently in such systems). A question the computer may need to answer, for example, is "*Who was the first prime minister of Canada?*" In such cases, knowing a category (hypernym) relationship between "*Sir John A. Macdonald*" and "*prime minister*" would mean a better chance of producing a correct response.

Another area where the results of the proposed method could be useful is that of Named Entity Classification. Named Entity Classification is the problem of giving Named Entities (such as *University of Alberta*) categories. Typical categories include "Person" and "Company". Named Entity Classification may also be improved by the ability to classify Named Entities more specifically than by the usual techniques. For example, given the named entity *University of Alberta*, it is more specific (and useful) to label it as a "university" than it is to label it as an "organization". For Named Entity Classification, the category given by the method proposed in this work can be more specific than that given in conventional techniques.

The organization of this work is as follows: Chapter 2 contains the background materials and information needed for the understanding of the new algorithm. Chapter 3 describes other techniques that have been used in the automatic gathering of semantic relations. Chapter 4 gives the details of the new algorithm. Chapter 5 consists of the experiments performed and the results produced by the algorithm. Finally, Chapter 6 sums up the work and proposes avenues for future research.

Chapter 2

Background

This chapter defines some important background terms and details the tools used in the development and testing of the new algorithm.

2.1 Definition of Common Terms

- **Aggregate Data**: Aggregate data refers the combined data of all the terms in a given noun cluster.
- **Cluster**: A cluster refers to a group of similar expressions, as gathered by an automatic clustering method.
- **Cluster Labeling**: The process of assigning hypernym-hyponym relationships to a group of expressions will be referred to as Cluster Labeling.
- **Dependency Relation**: A syntactic relation between parts of a sentence is referred to as a dependency relation. An example would be an appositive relation, such as "Smith, the chairman" which would be described as a *N:appo:N* relation from the point of view of *Smith*, and a *-N:appo:N* relation from the point of view of *chairman*.
- **Holonym**: A holonym is a term used by WordNet [8] to describe the part-of relationship. Here, one expression is a holonym of another expression in the case that it is a part of another. An example of a holonym would be the relationship between *wheel* and *bicycle* (*wheel* being a part of a *bicycle*).

- **Hypernym**: A hypernym is defined as the relation between an expression and an expression that is superordinate to it [6]. Suppose there are two expressions A and B. If everything that applies to B is true of A, then B is a hypernym of A. An example would be the relationship between *doll* and *toy*. *Toy* is a more general class than *doll*, so *toy* is a hypernym of *doll*.

- **Hyponym**: A hyponym is just the opposite of hypernym.: thus *doll* is a hyponym of *toy*.

- **Named entity**: A Named Entity is an expression that is described by a proper name. For example, *John Smith* would be a Named Entity, as would *The University of Alberta*. Named Entity Classification is normally performed as a process following the identification of the named entities in text, but in this project, named entities are not the only expressions being worked with. For this reason anything that would not fall under the definition of named entity will be referred to as a regular entity.

- **Precision**: Precision is defined as the percentage of answers returned that are correct [6].

- **Recall**: Recall is defined as the number of correct answers returned divided by the number of correct answers possible [6]. For this problem, it is not possible to determine all of the possible answers, so recall will refer to the number of clusters that have a correct label returned.

- **Regular Entity**: A Regular Entity is an expression that is not a Named Entity. So, an expression like *bookstore* would be a regular entity.

- **Unsupervised Learning**: A learning method is referred to as being unsupervised if it does not use the answers to the problem while training. Conversely, a learning method that does use the answers (such as having annotated data, where the correct results are given) is referred to as supervised ([12], p. 528).

2.2 Tools

The following pre-built tools were used in the development and testing of our algorithm.

2.2.1 Clustering by Committee

To be referred to as CBC for the rest of this work, the Clustering by Committee method of automatic clustering developed by Patrick Pantel [9] is the technique that produces the clusters worked with in this thesis. The basic idea behind this clustering method is to use a committee in each cluster to attract other terms similar to it. Selecting a small number of very strong terms for the committee allows the concepts to have less interference from expressions that have multiple concepts that they should be placed in.

In [9], the algorithm for producing these clusters is described as having three stages. In the first stage, the top-k most similar elements for each element are calculated, forming a similarity matrix. The second step produces the committee clusters, which are tight clusters. This is a recursive procedure, that repeatedly calculated the tightest clusters and what elements would be closest to them. At the end of stage two, the committees have been chosen. Lastly in phase three, the clusters are expanded to their final size as elements are added to the clusters to which they have high enough similarity scores (each element can be assigned to more than one cluster).

The dependency data and the cluster data were extracted with the use of LaTaT database tools (Lin [7]) before further processing.

2.2.2 WordNet and QueryData

Aside from the clusters, there were also some other tools used in the evaluation of this method. Because WordNet is so often used as a resource, some experiments here incorporated it. WordNet [8] is a hand-built lexical database that has been worked on and improved over a considerable period of time. WordNet version 2.0 is the version used for testing in this work. To use the WordNet information, the use of another package, known as QueryData [11],

was incorporated. This package allows easy use of the information stored in WordNet.

Chapter 3

Related Work

This chapter will discuss four different approaches that have been used to automatically produce hypernym relations. They are organized here based on the type of data that they start with. The first type is those methods that are based off of machine readable dictionaries. Then there are methods that extract the relations directly from text. Also present are methods meant for Named Entity Classification, and finally methods that start with clusters and classify them with hypernym relations.

3.1 Dictionary Based Methods

In work by Chodorow et al [2], the data available in machine readable dictionaries is used to extract hypernyms with the goal of producing a semantic network. The idea is to extract hypernyms from the definitions of the nouns. Typically, the head of the definition phrase was taken to be the hypernym of the word being defined. For example, if a definition of golden retriever is a golden-haired dog, then dog would be taken as the hypernym of golden retriever. Some simple heuristic rules are used to extract the head of the definition, and then with substantial human involvement the information could be worked into a hierarchical structure.

A later paper by Ide and Veronis [5], examines the progress made with these methods, with the conclusion that a dictionary by itself is not enough to automatically produce the correct results. This conclusion is attributed to the varied structure in the dictionary definitions, missing information, and the

fact that different dictionaries produce markedly different hierarchies.

3.2 Text Mining Methods

Around the time that the dictionary techniques were becoming less popular, computers were able to start processing much larger amounts of data in a tractable amount of time. Before the clustering techniques, but after the dictionary techniques, there is the following work on a plain text corpus.

In a paper by Hearst [4], hypernym relations were extracted directly from text by the means of syntactic patterns. The first stage of this technique was to produce a list of patterns. Patterns were found by using a known relation to track down possible patterns. For instance, using *toy, doll* as a known relation, patterns such as *dolls, tops and other toys* is easier to notice. In all, six patterns were defined. Only one of these was used in their testing (corresponding to a construction similar to *A toy, such as a doll...*). Hypernym/hyponyms were extracted from encyclopedia text and evaluated against the structure of WordNet. Through this process, 152 relations were found, of which 106 had both terms in WordNet. For example, a relation might have one term in WordNet, and one that was not. These cases are not as easily evaluated. Of the 106 possible relations where both terms are present in WordNet, 61 of the possible relations already existed.

3.3 Named Entity Methods

In a paper by Fleischman and Hovy [3], the goal of getting specific classifications for Named Entities was pursued. They worked on classifying names into a few categories, such as politician, businessman and entertainer (eight categories in all). Fleischman and Hovy produced a feature based system for classification. The features used in this case were previous and following n-grams, topic features (a calculation of how much a some words are likely to indicate category), and WordNet features. With these features they used several methods to train classifiers, including decision trees (C4.5), neural nets, a support vector machine, and Naive Bayes. As all the learning methods are

supervised techniques, training these methods required a correctly classified training set. The set of correctly classified names (estimated to be 99% correct) was produced by a semi-automatic process combined with manual work. The best results are with the decision tree method, which with all the features achieves 70.4% accuracy.

3.4 Cluster-based Methods

Caraballo [1] uses a clustering based method to produce hypernym/hyponym information as well. The goal of this work is to build a semantic lexicon by automatically creating a hierarchy from a body of text. The hierarchy consists of noun clusters grouped under their hypernyms. The hypernyms are added during the latter part of the construction. First, groups of similar nouns are gathered. It is interesting to note that only a few syntactic relationships are included in this process, namely the conjunctions and appositives. It is from these relationships that the clustered nouns are extracted. Clustering the nouns with a bottom-up method produces the general hierarchy, as similar nouns are placed under the same parent in a binary tree.

The hypernyms are subsequently applied to the structure. The hypernyms were obtained by gathering the details of a specific sort of conjunction, namely the ones that have the word "other" in them. For example, "cars, trucks and other vehicles" would indicate that "vehicle" was a hypernym of "car" and of "truck". For each of the clusters in the hierarchy, up to three possible hypernyms were assigned.

The results in this study are evaluated by judges, and the range of correct hypernyms ranged from 33% to 60%. The evaluation rates the hypernyms from strictly correct to leniently correct, so 33% are strictly correct while 60% are leniently correct. Strictly correct means that the majority of the judges agreed, while leniently correct means that at least one judge thought it was correct.

One recent set of results comes from Pantel and Ravichandran [10]. The clustering by committee algorithm produces the clusters used in this method.

This method consists of three stages. Stage one requires calculating some vectors for each word in the clusters, namely a frequency count vector, and a mutual information vector that gets discounted to reduce the effect of data sparseness. The second phase produces a committee for each cluster. The goal is to isolate the most representative members of the cluster. Phase three forms a signature for each cluster, derived from the feature vectors of the committee members.

The method to choose a possible hypernym restricts the possible label words to a smaller set of features. Through human evaluation, four particular features were found to be the most important. These four relations are: *N:appo:N, -N:subj:N, -N:such as:N and -N:like:N.* From these four relations and mutual information scores, the scores for possible labels are calculated.

The results were evaluated by human judges. The first answer that was given by their method is reported to be correct 72% of the time.

Chapter 4

The Technique

My new algorithm for cluster labeling is an unsupervised learning method. The reason for developing an unsupervised technique is to avoid having to annotate the data, which is time consuming and difficult.

After examining the data available, it is quickly noticed that possible labels for clusters are often present in the aggregate data. Consider Figure 4.1. In this sample, the cluster consists of horse race names. In the excerpt from the dependency data, *race* is displayed in bold, and clearly it is a good label for this cluster (*event* is also a reasonable choice). This trend continues throughout the clusters and their dependency data, leaving the problem of how to extract the good labels. However, there is a huge number of feature words for some features, and many are not useful.

Cluster	Features and possible terms
Preakness Stakes	-N:before:N 23
Preakness	day 19
Belmont Stakes	start 2
Travers	**race 2**
Santa Anita Derby	-N:subj:N 80
Kentucky Derby	**race 51**
Florida Derby	run 7
	goal 7
	event 8
	victory 3
	start 2
	history 2

Figure 4.1: Sample dependency data for a noun group

4.1 Motivation

This algorithm is developed to learn weights for each feature in order to pick out the good labels from the rest of the data. It is not merely a case of picking the most common feature or the most frequently occurring feature word. Instead, the feature scores need to reflect the importance of the feature, so that good labels are found in features with high scores.

To this end, the iterative algorithm described here was developed. Each iteration of the algorithm redistributes the feature scores to better represent their values in relation to possible labels. As a feature score increases (or decreases) through iterations, so do the label scores of the feature words associated with that feature. As a result, the feature scores of important features are amplified by the presence of good labels in that feature, while unimportant features are given low scores. Consequently, the good labels are more likely to be present in features with high scores than in features with low scores.

4.2 Terminology

The features for this method are the dependency relations as described in Section 2.1: the syntactic relations between parts of a sentence. The feature set consists of all such dependency relations and is denoted by F, and an instance of a feature will be referred to as f. All of the features used in this process are represented by expressions like *N:appo:N*. For a graphic example, see Figure 4.2, where the basics of a set of data are demonstrated. Each cluster may or may not have a certain feature associated with it, depending on whether any words in the cluster have terms in that dependency relation. The list of specific features found in the data can be found in Appendix B.

The following definitions are important to the explanation of the algorithm in this chapter.

Feature Score: The feature score is a numerical score assigned to a feature. The value of such a score will be represented as *FS(f)*. The scores range from zero to a possible maximum of the number of features; however a differ-

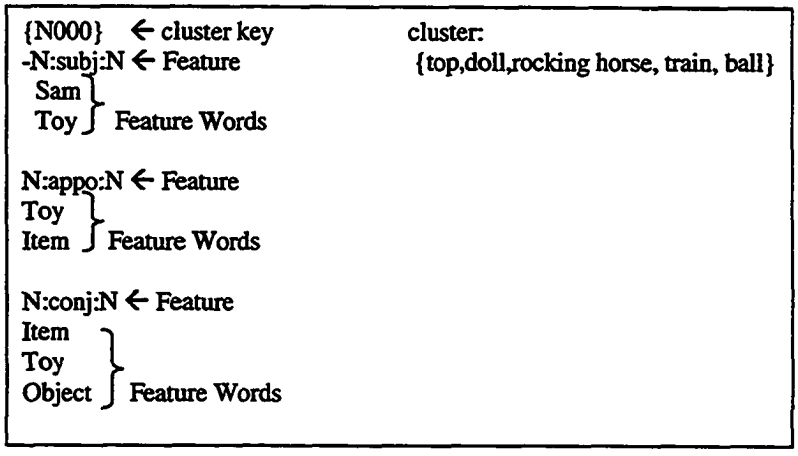

Figure 4.2: Examples of Terms

ent choice of normalization method can be used to produce other ranges. A zero score indicates an unimportant feature, while a high number denotes an important feature. The feature score for a given feature is produced during the training stage of the algorithm.

Feature Word: A feature word is a term that has been listed with a feature in the data set. A term *a* is said to be a feature word of a feature *f* if and only if *a* is listed as an instance of the feature *f*. The following relation is defined to indicate the status of a feature word with respect to a feature: *C(f,*a*)* is true if and only if the feature *f* has *a* as a feature word. In Figure 4.2 with the small data example, *toy*, *item*, *Sam*, and *object* are all feature words. Also, the relation *C(N:Subj:N,Sam)* is true.

A feature word is associated with a feature if it has occurred in the dependency relation indicated by the feature. For example, in Figure 4.2, *toy* occurs as a feature word of the feature *N:appo:N*. This could indicate that a phrase such as "Sam threw the ball, a toy." has occurred, giving *toy* an appositive relationship to the cluster.

Feature Vector: A feature vector is the list of features in which a given feature word occurs. It is used frequently in the implementation. For example, in Figure 4.2, the feature word *toy* has a feature vector of (*-N:subj:N, N:appo:N, N:conj:N*), while the feature word *item* has a feature vector of (*N:appo:N, N:conj:N*). Therefore, a feature vector of *a* is just the list of fea-

tures for which *C(f,a)* is true.

Label: A label is a term that has been selected from the set of feature words to categorize a cluster of noun terms. This is frequently a single noun, but multi-word terms do sometimes appear in the data.

Label Score: A label score is the total rating a possible label has been given. It is defined by the following formula:

$$LS(a) = \sum_{f \in F} FS(f)C(f,a)$$

where *C(f,a)*=1 if true and 0 if false. The formula produces a summation of the feature scores (*FS(f)*) over all the features in the feature vector of *a*. As an example, in Figure 4.2, to find the label score of *item*, the feature scores of *N:appo:N* and *N:conj:N* would be added together.

Normalization: Normalization is the process of bringing the feature scores of all the features back to an average of one. Normalization is applied as a multiplier to the feature score of each feature. The formula below shows how the multiplier is calculated: *FS'* represents the temporary feature score obtained during execution, and *b* represents the normalization factor applied to a feature score.

$$b = \frac{|F|}{\sum_{f \in F} FS'(f)}$$

The sum over the temporary feature scores will never be 0, since this would require all feature scores on the last iteration to be 0 as well (the feature scores start at 1 and are multiplied be a number larger than 0 every iteration, so this cannot happen without initializing differently).

4.3 Data Preparation

The clusters used for this research are a subset of the clusters produced by the Clustering by Committee method of Pantel [9], as described in Chapter 2. The clusters originally consisted of noun, adjective, and verb clusters. Only the noun clusters were used, to narrow the problem down. The aggregate data of the clusters was used, which consisted of the combined dependency data retrieved from the corpus.

The data used in the learning method was available through the use of a system called LaTaT (see Section 2.2.1). To extract the information needed, the LaTaT database needed to be searched. Details of the script are included in Appendix C. The data returned from these queries was compiled into a data file, and information that is not used by the algorithm was filtered out. Because the goal was to produce a label for a group of nouns, labels other than nouns do not fit logically into the problem (that is, verbs and other grammatical types are not reasonable labels for a cluster of nouns). The elimination of verb and other grammatical types was supported by early attempts that selected highly inappropriate verb labels for clusters. Additional information from the LaTaT queries that was not used in the learning method was also removed from the data file.

There are several files that hold the information vital to the method. The feature file contains the list of features, and their current feature scores. The data file contains the dependency relation information separated into a different section for each cluster.

4.4 System Overview

Initially, I experimented with a probability model for the labeling of clusters. Several variations were tried, but either proved unsatisfying, or failed to perform adequately after implementation. An earlier version before the final implementation was a method that attributed a probability to each feature word. This version started with all labels having the same probability, and changed them through iterations. This model did not make good theoretical sense, as there certainly could be more than one good label for a cluster, and in practice it tended to prefer features that had many feature words, rather than giving all features a chance. Finally, a scoring model was chosen, which is described below.

The input to the training method is the dependency data of the clusters, and the list of features. The expressions in the clusters themselves are not used as input to the process. The input to the labeling method is the dependency

data along with the newly produced feature scores.

The output of the training process is a feature score for each feature in the feature set (f in F). The output of the end process is a ranked list of possible labels for each cluster. The list can be returned in a top-k or a score threshold fashion.

4.5 Training the Model

With the data established, the main procedure can be developed. Figure 4.3 shows the pseudo-code of the unsupervised training process.

The training method is an iterative approach. Initially, the scores assigned to all the different features are the same. All feature scores being equal represents the state of not knowing anything about the data yet. The feature score initialization step corresponds to the first loop (lines 1 and 2) in the figure 4.3, where *FS(f)* is set to one for all features.

The main loop (lines 3 through 14) begins the actual iteration through the learning process. During an iteration, the label scores for each feature word are calculated from the current feature scores, produced by summing all the feature scores of the features that the feature word is in (that is on lines 7 and 8 in the pseudo-code, where *LS(a)* is calculated). For each feature, a temporary score is maintained (denoted by *FS'(f)*), with the label scores of each feature word in that feature summed into it. The temporary scores are calculated on lines 9 and 10 of the pseudo-code.

After the temporary scores, δ is calculated (line 11 of the pseudo-code), which is a variable used to keep track of convergence. In theory, a threshold might be set for convergence, but in practice the values for the feature scores have always converged, so here the algorithm stops when the values of the feature scores stop changing.

After each cluster has had the values calculated, the feature scores of each feature are replaced with the normalized temporary score. This process happens after each iteration, on lines 12 and 13. This iterative process is continued until the feature scores converge.

1) **For** each f in F **do**
2) $FS(f) := 1$
end for
3) **Repeat**
4) **For** each f in F **do**
5) $FS'(f) := 0$
end for
6) **For** each Cluster in the training set **do**
7) **For** each feature word a in the cluster **do**
8) $$LS(a) := \sum_{\substack{f \in F \\ C(f,a)}} FS(f)$$
end for
9) **For** each f in F **do**
10) $$FS'(f) := FS'(f) + \sum_{C(f,a)} LS(a)$$
end for
end for
11) $\delta := \sum_{f \in F} |FS'(f)b - FS(f)|$
12) **For** each f in F **do**
13) $$FS(f) := \frac{|F|}{\sum_{f \in F} FS'(f)} FS'(f)$$
end for
14) **Until** $\delta = 0$
15) **For** each Cluster in the testing set **do**
16) Label:= $\underset{a}{Max}\{LS(a) := \sum_{\substack{f \in F \\ C(f,a)}} FS(f)\}$
17) **Print** Label
end for

Figure 4.3: The Training Algorithm

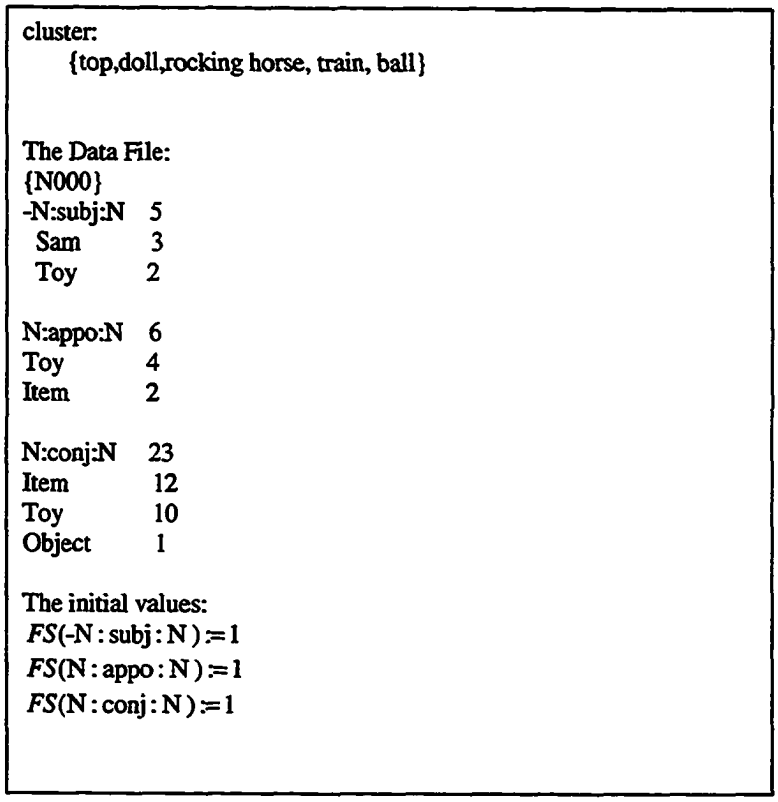

Figure 4.4: The Feature Score Training Example: Initialization

For an example of how the procedure executes, we return to the example cluster introduced earlier. Figure 4.4 shows what the initialization step of the algorithm will produce on this toy example.

After initialization, the iterative calculations begin. In our example, there is only one cluster, three features, and four feature words to deal with. For an example of the calculation, consider the feature word *item*. It occurs in two of the features, so it has a feature vector [*N:appo:N, N:conj:N*]. The value of *LS(item)* during this iteration is then 2, since it gets the sum of the two feature scores, which were both initialized to 1. The other feature words are calculated in the same way. After the first pass through the cluster information, the FS' scores are displayed in Table 4.1.

After the temporary scores have been calculated, they get normalized. The normalizing factor, b, as presented earlier, will have a value of $\frac{1}{5}$. Thus the new scores of the features in the next iteration are $\frac{4}{5}$, 1 and $\frac{6}{5}$.

Since the feature scores have not converged after one iteration, the process

FS' values	
FS'(-N:subj:N)	4
FS'(N:appo:N)	5
FS'(N:conj:N)	6

Table 4.1: The FS'(f) scores after the loop

LS(a) values	
Sam	0.72
toy	3.0
item	2.28
object	1.26

Table 4.2: The LS(a) scores after convergence

continues. After several iterations, the scores stop changing, and the algorithm produces final values for the feature scores as given in Figure 4.5.

The input to the actual labeling step is the cluster dependency information along with the previously calculated feature scores.

The labeling is assigned by summing the feature scores that each word occurs in. Also, if two feature words happen to have the same score, the more frequently occurring one will be listed higher on the list (although this does not change the score itself). So considering again the toy cluster example, from the final value figure 4.5, it is clear that the first label chosen for the cluster given by this method would be *toy*, as it has a much higher score than the other possible labels. The scores of all the possible labels are shown in Table 4.2.

4.6 Variations on the Basic Method

For the purposes of evaluation and potential improvement of the method, several variations on the basic training method are used. In the following section, the variations are described, including adjustments to the training method, the baseline to be used, some heuristics, and the use of WordNet.

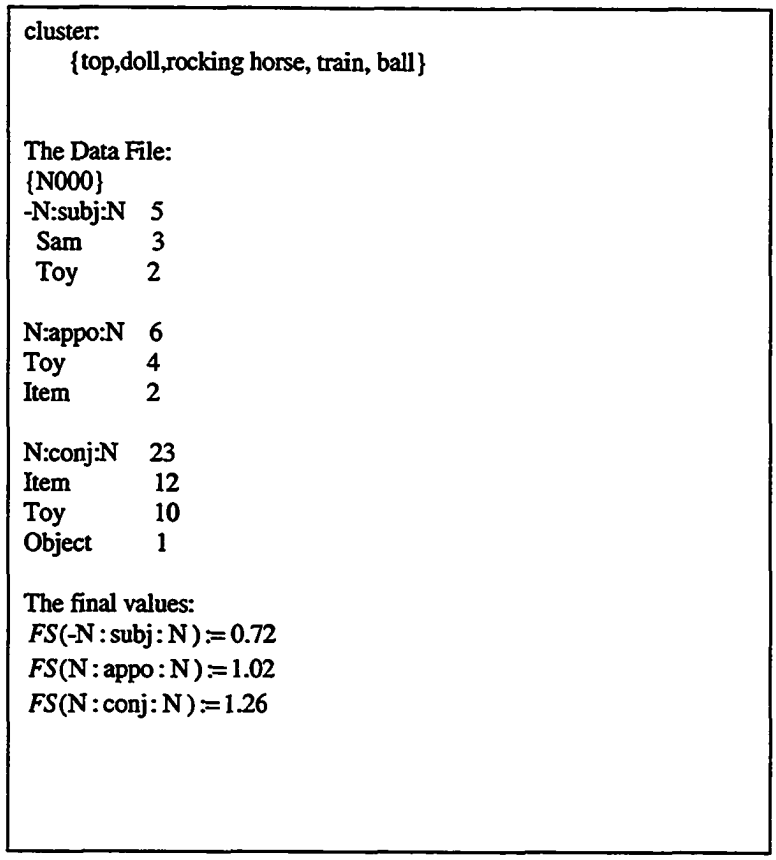

Figure 4.5: The Feature Score Training Example: Final Values

4.6.1 Frequency Filtering

How often a feature term occurred was not taken into account in the base training technique. The process of frequency filtering attempts to remedy that by using the frequency information (frequency of the feature words, not the cluster terms) to filter out a part of the training data. Specifically, if a given feature word has a frequency less than a certain threshold, that feature word would be ignored in training. The threshold is in respect to the number of occurrences a feature word has with respect to the number of times a feature occurred. Frequency filtering can also be applied during the label selection step.

As an example, consider again Figure 4.4. If a filtering rate of 10% is applied to this cluster, then any feature word that does not occur at least 10% of the time in a feature would be ignored. The cluster has one such feature word, *object*. *Object* does not get considered because it occurs only 1 time out of 23 occurrences of the Feature *N:Conj:N*. If frequency filtering is applied during training, then *object* would not be used in the training process, while if filtering is applied during label selection, it would not be considered as being a possible label (filtering during the labeling step turns out not to be very useful).

4.6.2 Baseline Method

A baseline method had to be chosen for comparison with the results. Since the algorithm described above assigns variable scores to the various features, the baseline method used for the baseline to assigns all the features equal scores. The baseline has the effect of making the feature words that occurred in the highest number of features get the highest score. It is these feature words that end up being chosen as labels. The baseline then, is equivalent to stopping the learning algorithm before the first iteration, so it is really a base version of the learning algorithm.

Answer Distribution Features	Features of Pantel and Ravichandran
-N:subj:N -N:conj:N -N:appo:N -N:nn:N N:subj:N N:conj:N N:appo:N N:nn:N	N:appo:N -N:subj:N -N:such as:N -N:like:N

Figure 4.6: The Feature Sets

4.6.3 Heuristic methods on the feature data

The baseline is one method used for comparison with the learning method, but some others were also considered. A fairly simple additional technique was to consider only parts of the feature set. This restricted the training data and the possible labels. The heuristic methods were able to avoid some irrelevant or less useful features.

Considering only part of the feature set resulted in a simple heuristic, where a subset of the total features was chosen, and each feature in that set was weighted equally for the purposes of selecting labels. Another variation on this is to train the main method on a reduced feature set.

Two specific feature groupings were used for feature restriction purposes. The actual grouping used are shown in figure 4.6. In the first grouping of features(the one termed Answer Distribution Features, to be called ADF from now on), the members were tailored to the development set. In the development set, the hand-selected answers resulted in a distribution of features in which they occurred. For example, if a hand-selected label occurred in the features (*-N:conj:N*) and *-N:Subj:N*, each of those features get a count of one. The full list of features associated with answers in the development set is shown in Figure 4.7. If a feature has no occurrences of answers, it is not on the list. The first heuristic set has the most common feature pairs on this list included, where a pair is a feature and its inverse relation (*N:conj:N* and *-N:conj:N* being an example). The second grouping of features, termed the

N:conj:N	120	N:of:N	7	-N:near:N	2
-N:conj:N	117	N:in:N	6	N:into:N	2
N:appo:N	92	-N:person:N	6	-N:subj-in:N	1
-N:nn:N	79	N:as:N	6	-N:over:N	1
-N:appo:N	74	-N:at:N	5	-N:by:N	1
-N:subj:N	74	N:to:N	5	-N:below:N	1
N:nn:N	49	-N:with:N	4	-N:among:N	1
-N:of:N	48	N:for:N	4	-N:subj-on:N	1
N:subj:N	43	-N:like:N	4	-N:through:N	1
-N:gen:N	14	N:gen:N	3	N:over:N	1
-N:in:N	13	N:on:N	3	N:UNDER:N	1
-N:such as:N	12	N:at:N	3	N:before:N	1
-N:for:N	11	N:from:N	2	N:through:N	1
-N:on:N	9	-N:into:N	2	N:subj-on:N	1
-N:as:N	8	-N:than:N	2	-N:after:N	1
-N:from:N	7	N:with:N	2	-N:since:N	1
-N:to:N	7	-N:before:N	2		

Figure 4.7: Answer to Feature Distribution

Features of Pantel and Ravichandran(PRF), are the features listed in [10] as the top features found in that labeling method(this method is described in detail in Section 3.4).

4.6.4 Named Entities

The noun clusters are heterogenous (for a sample of the one hundred clusters used in the test set, see Appendix D). One useful property of the data is that some clusters are primarily named entities and that others are primarily regular entities. As the named/regular entity combinations could have had a large impact on results, a method was constructed to separate them to compare the differences.

The procedure used was to classify a cluster as either a named or regular cluster. A cluster was considered a named entity cluster if a majority of the terms were named entities, and a regular entity cluster otherwise. So from the example given in Figure 4.4, the cluster contained the terms *top*, *doll*, *rocking horse*, *train* and *ball*, none of which were named entities, so that cluster would be considered a regular entity cluster. A cluster such as (*John Smith*, *John Doe*, *Mr. White*) though, would be considered a named entity cluster. The named entities were sorted from the regular entities by considering the capitalization of the terms in the cluster. The separation of the regular clusters from the

named entity clusters is another option in training and labeling, that allows calculations of different feature scores on both types.

4.6.5 Making Use of WordNet

The WordNet resource being widely used, often in comparison, made it a logical choice for comparison with the learning method. Comparison to WordNet was developed by considering the hypernym entries for all the terms in each cluster. Then all of those results had their hypernym entries included, and so on up to the top of the hierarchy. The root of this chain was filtered out, since categories such as *artifact* are very general, but come up often. The purpose of this process was to get the maximum number of possible terms, so all the ancestors(other than the root) of the cluster term in WordNet were selected as possible labels. The reason for collecting the maximum number of terms was to provide the best comparison possible for recall of labels found in the training method. Terms generated from WordNet can be filtered and restricted the same way as in the training data.

4.6.6 Combining Feature Score and WordNet

After using the WordNet information for comparison purposes, another use was found. WordNet and the training method proved to have many good labels, but both methods also proved to have a poor precision rate. As an addition to the other variations tested, in order to improve precision, a combination of the training method and the WordNet method was developed.

In order to combine the methods, the lists produced from each method were intersected to produce a new much shorter list. Of course a down-side is that, if the WordNet method did not have entries for that cluster, the intersection would be empty. WordNet data does not always overlap with the training data either, so when WordNet does have a choice of labels, the intersection can still be empty.

Chapter 5

Experimental Setup and Results

The testing of the unsupervised learning method consists of many experiments. The experiments are detailed here, covering the methods and variations discussed in Chapter 4.

5.1 Experimental Setup

Two measures are used throughout this chapter to measure performance on the task of cluster labeling. The measures are precision and recall. The recall is calculated differently than the typical formula because the task of labeling the clusters does not have a deterministic method for defining the correct answers. As a result, recall is instead determined as the percentage of clusters that were assigned a possible label that occurs in the answer key for that cluster. The precision is just as reviewed in Chapter 2, the percentage of answers returned that are judged correct.

As an example, consider a cluster containing: *top, doll, rocking horse, train, ball.* If the answer key for this cluster consists of *toy, object*, and the possible labels given by the system consist of (1. *item* 2. *toy*), then the recall would be 100% (for getting an answer from the key) and the precision would be 50% (as half the possible labels returned are in the key).

For testing, the clusters are also divided into Named and Regular Entities (as detailed in Chapter 4). For the development set, the result is a group of 60 Regular Clusters and 40 Named Clusters. The clusters from the development set are manually checked to ensure that they are categorized correctly and no

errors were found. The Test Set ends up with a group of 34 Named Clusters and 66 Regular Clusters.

Another important thing to note is that when the training method is tested on Regular and Named Entities separately, it trains on the two types separately as well (so when testing on the Named Entities, it trains only on the dependency data from the Named Entity Clusters).

The development set consists of the first one hundred clusters represented in the data file. The clusters as well as their data file are used in the training and testing of the development set. Whenever development set results are mentioned, this is the data that is used. This set was chosen and separated from the total set of data to avoid tailoring the method to results in final testing.

The test set also contains one hundred clusters. In this case however, the one hundred clusters are selected according to their key number. The first one hundred clusters (only of those not already in the development set) with a key number divisible by ten are chosen as the set of clusters for the test set. The clusters and their data are held out, and not used in either the development phase or the training stage of the testing. There is no overlap between the development set and the test set.

The term "training set" in this research refers to the data used in the final training of the learning process. The final training involves a data file with the dependency(feature) information of all of the noun clusters available, except for clusters that were in the test set. Therefore the training set includes the clusters from the development set, but no data from the test set.

For the purpose of training during development, the dependency data from the 100 clusters in the Development Set is used. All the remaining clusters and feature data are held out during the development testing.

An answer key is manually generated for the development set. The process involves going through each cluster, and choosing labels for the clusters from the dependency data. As a result, the answer key for the development set has labels that occur in the data. The goal of selecting answers for the key in this way was to produce an answer key with the answers that could be found in

the data, rather than answers that the method may not be able to find at all. Some clusters have no appropriate label in the data, and remain unlabeled in this answer key. The answer key was produced to try get the most specific terms possible (that is to choose a label that would be close to it in an ontology hierarchy). General terms, such as *person*, or *location*, are not set as answers in the key unless there are no more specific terms to fit the cluster. As a better example, plenty of the Named Entity clusters could be labeled with *people* or *person*, but this is not the best label in most cases. A cluster of hockey players is much better labeled with *hockey player* than with *people*. After the clusters in the development set were hand-labeled, the list had between zero and five answers for each cluster in the set. Out of the one hundred clusters, five did not have an appropriate label within the dependency data. As a result, the best recall that the computer can hope for on the development set was 95%.

Answers are returned either as a list of a top-k list of labels, or by a score threshold. The score threshold will return answers with a score that is higher than the threshold value, which can result in no answer being returned on some occasions. The top-k list always returns k answers unless the data for a cluster was so sparse that there was not k possible labels to choose from. Although returning by threshold looks like a good way to return cluster label lists, in practice it gives inconsistent numbers of labels for each cluster, due to the sparseness of some the data. Some clusters had many more features and possible labels than others.

5.2 Human Performance Experiment

To form the answer key to the Test Set, and to examine human performance on the task of classifying clusters, a study was conducted. Participants were asked to classify the clusters in the test set with the best labels they could come up with. The clusters were labeled without any outside influences (for example, looking up all the terms in a cluster was not permitted). There was the option of indicating that a cluster has no labels. Needless to say, the average time for a human to complete the labeling of the clusters is considerably longer than it

takes the computer system. Altogether, eight people participated in the study.

The instructions given to the participants are as follows:

Label the groups of terms given with appropriate categories(s). You may have more than one classification. You may not use any outside aids to choose the classification (you cannot look up any terms, or ask anyone about terms). For each cluster, fill in the corresponding line on the response page with your labels. Note: terms that are entirely uppercase have no particular significance. If a cluster has no category, please check the no answer box.

Example: Qinghua University, Beijing University, Nanjing University, Shanghai Jiaotong University, University of China, Zhejiang University, Nankai University, University of Hong Kong, Chinese Academy of Sciences, Qinghua, Chinese Academy of Social Sciences, Chinese University of Hong Kong, Peking University, National Library, Museum of Chinese History, University of California, Beida, Center for Strategic, Tulane University, Cass, Museum of Natural History, preparatory school, Association for Friendship, John F. Kennedy School of Government, Bell Laboratories, State Administration of Cultural Heritage, CAS, Chinese Academy of Engineering, Lawrence Livermore, Bank of Israel, yeshiva

Given the terms above, you might label this group: University, academy, school. If you were unaware that these terms were institutions, you might label this group: Location.

The list of labels given by the human participants in aggregate form is in Appendix E. The average number of labels that each participant put to each cluster is .96 labels/cluster. Agreement between labels from different testers is at 42.8%. Average agreement is calculated as the average number of participants who agree on a label that is the result for the highest number of participants. Therefore this statistic also includes the cases where the most common answer was the response "no answer" (there is several instances of this). There is no instances where all the participants assigned the same label to a cluster, as well as several where no two participants have the same label. Obviously the clusters vary in how easy they are to label, and how many acceptable labels there may be for each cluster. Many of the participants

indicated that the labeling was very difficult.

5.3 Development Set Results

The following sections describe the results of the experiments on the development set. The graphs showing results throughout the rest of this chapter have data points for different numbers of possible labels returned, with the exception of the WordNet results. All of the experiment run on the training method show data points for the following numbers of possible labels returned: (1, 2, 4, 8, 10, 14, 18, 20). The intersection experiments have data points for 1 and 5 labels returned, and WordNet has a single point for all labels returned.

5.3.1 Baseline Experiment

Stage One of testing involves setting a baseline condition to compare the other methods against. WordNet could be considered as a baseline, but is not easy to compare to because it lacks many of the terms in the clusters. Instead, the baseline is a feature score distribution where each feature is given the same score (i.e. the initialization state of the system).

The baseline performed poorly on the Development Set. Out of the 100 clusters, the first answer given by the baseline was among the possible answers for 24 of the clusters. Even with the number of possible answers set to 4, the baseline only manages to give a label from the answer key to 52 clusters. As a result, the precision rating for the baseline is very low here, only 16.8% with four returns. In Figure 5.1, the baseline recall vs. precision is shown with the other results on the development set. The graph shows that precision is low for most of the experiments, and drops rapidly as more answers are returned. The rapid drop is the result of the number of answers returned ending up very quickly bigger than the possible number of correct responses listed on the answer key. In this graph, as in all the graphs in this chapter, the first data point (where recall is low) represent the precision and recall with 1 label returned, and it moves up through to a total of 20 labels returned. The baseline is well below all the other methods tested here, aside from the PRF

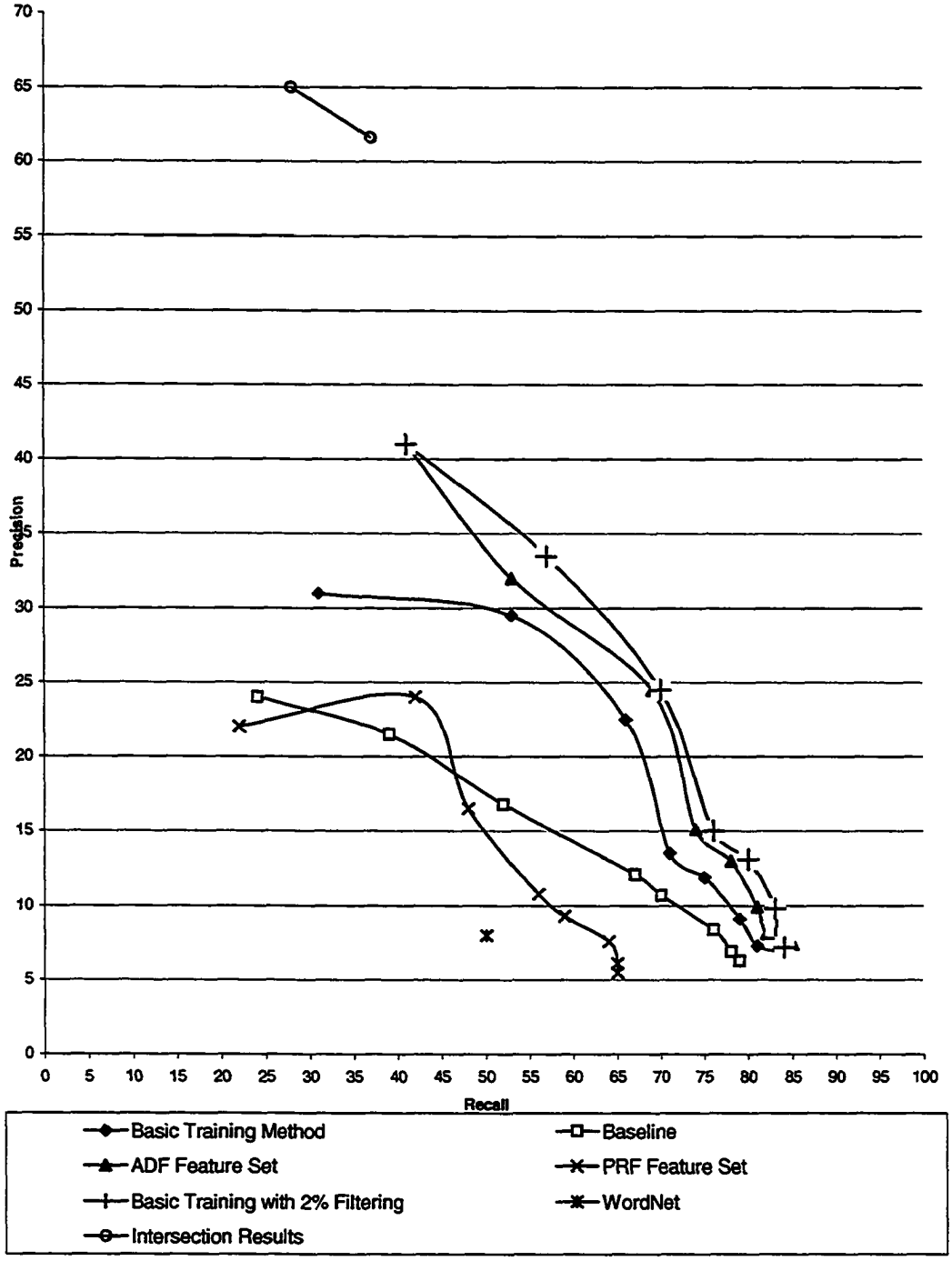

Figure 5.1: Results of Experiments on the Development Set

set, which behaves rather strangely.

5.3.2 Heuristic Experiment

In the Heuristic Experiment the feature set methods are also compared to the unsupervised learning method set out in Chapter 4. As mentioned earlier, there are two main sets that are tested. The results on the Development Set are given in Figure 5.1. The Answer Distribution Feature Set (referred to as ADF, and described in Chapter 4) already has a large improvement when compared to the baseline. It starts off with nearly double the correct answers. When the ADF Set returns a single possible answer for the development clusters, it gets correct responses in 41 of the 100 clusters. After it is returning four possible labels for each cluster, the success rate increases to 69 labels correct. The results of the ADF set ranked against the other methods and variations are given in Figure 5.1. Given that the ADF set was created to have as many correct labels as possible, it is unsurprising that it does so well against the other methods here.

The experimental results of the Feature Set of Pantel and Ravichandran (PRF) look very different on the Development Set than the ADF Feature Set. The ADF Feature Set is based on where the answers in the answer key occurred in the data, while the PRF Feature Set is the set of features that was found most influential in the work of Pantel and Ravichandran [10] on labeling clusters (described in Section 3.4). The PRF Feature Set does not do nearly as well as the other set. The difference could be explained by the other heuristic over-fitting the Development Set (as would be expected given the way it was chosen). The results of the PRF Feature Set ranked against the other methods and variations are given in Figure 5.1. The PRF Feature Set actually performs worse than the baseline on the Development Set, getting successful labels in only 22 cases on 1 answer returned, and 48 on 4 returns.

5.3.3 Basic Method Experiment

Stage Three of the experiments involves performing the unsupervised training procedure, and then applying it to the Development Set. The first experiment

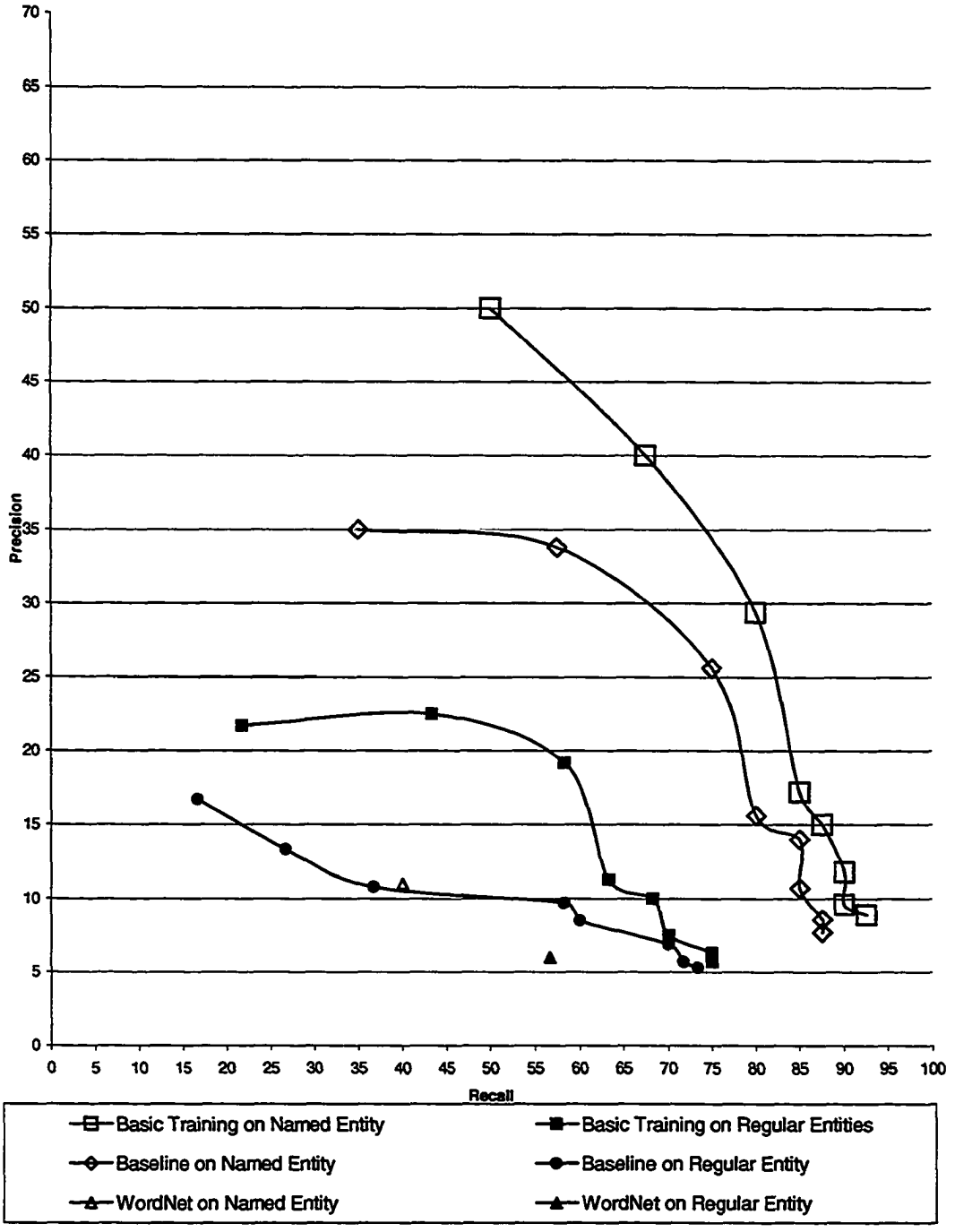

Figure 5.2: The Baseline and Basic Training Method Results on Named and Regular Entities on the Development Set

is the unmodified training algorithm results on the data from the Development Set. This means no feature filtering or other variants are used.

The results of the basic method on the Development Set ranked against the other methods and variations are given in Figure 5.1. The basic method performs considerably better than the baseline and the PRF set, since it is able to correctly label 33 clusters on the first returned label, and reached 66 correctly labeled clusters when 4 possible labels were returned. The basic training method produced feature scores that performed almost as well as the ADF feature set. Considering the algorithm has only trained on the data from 100 clusters here, it is quite effective.

Some more experiments were performed on the results of the basic training method. The learning method was next trained and tested on the Named Entity and on the Regular clusters separately. The goal of the experiments was to find out if there is a remarkable difference in what the training method was learning from the different types of cluster information. After running the basic training method on both types, the results certainly appear to indicate a difference in these two types of clusters.

The Named Entity set did very well when trained on the Named Entity part of the development set, and then applied to the Named Entity clusters to choose labels. It very quickly reaches a high recall rate, starting at a recall of 50% on 1 label returned, and getting all the way up to a recall of 80% on 4 labels returned. These results show a huge difference to the basic training results on the whole development set. The Regular Entity clusters suffer more under this method than the Named Entities did, with a starting recall of 21.7% success rate on the return of 1 label, moving up to a rate of 48% on the return of 4 labels. This may indicate that the labels for Named Entities are easier to learn than the Regular Entities. The results of these experiments are also shown against the rest of the development set experiments in Figure 5.1.

To examine the possible differences between the Named Entity clusters and the Regular clusters, the baseline is tested in the same fashion. Clearly there is some important differences between these two types of clusters, as the baseline has just as large a change between them as the training method. The baseline

does not perform as well as the training method in either case, but it also does much worse on the Regular Entities than on the Named Entities.

5.3.4 WordNet experiment

In this experiment, information from WordNet is compared to the other methods. WordNet does not have full coverage of all the clusters, but it is still useful to see how well it can do labeling the clusters, and later, being combined with the training method. It turns out that on the development set, WordNet can provide labels for 50 of the clusters, while having a precision of 8%. It is important to note here that this does not necessarily mean that WordNet was wrong on all the other clusters, as not all possible labels that WordNet comes up with are in the feature data, but merely that this is the extent that WordNet overlaps with answers in the answer key. It is also the case that for some clusters WordNet has no data present, so cannot attempt to label those clusters.

Here again there is a difference between the Named Entity and Regular Entity distributions. WordNet is more likely to have a valid answer for a Regular Entity than a Named one (at least in the Development Set). This is a positive sign, since the training method is already doing fairly well on the Named Entity clusters, but could use a bit of help on the Regular Entity ones. WordNet gives a recall of 40% with a precision of 11% on the Named Entity clusters, and a recall of 56.7% with a precision of 6% on the Regular clusters.

Since the method for getting the WordNet labels does not give scores the way the learning algorithm does, there is only one data point per WordNet experiment. The WordNet results on the graphs show that it provides high recall, but many labels that are not in the answer key (and may or may not be correct).

5.3.5 Frequency Filtering Experiment

The next experiment concerns modifications to the training method. It is often useful to consider that the noise in the data (here noise refers to parser errors, and/or tokenization errors that are present in the data file, as well

as feature words that occur very infrequently) is influencing the feature score training. One idea to combat this was the introduction of frequency filtering (described in Chapter 4). To see if this reduction of the noise in the data helps the training or labeling process some experiments are performed with various quantities of filtering.

It is be quickly noticed that the 2% filtering here has improved the initial training results obtained, by examining the graph in Figure 5.1. The initial recall with only 1 possible label returned has jumped a full 10%. This option is quite promising, and certainly seems to indicate that there is some noisy data that may be interfering.

5.3.6 Intersection Experiment

This section deals with the intersection method described in Chapter 4. The basic intersection result is the combination of twenty possible answers returned by the training method intersected with all the possible terms that WordNet provides. The results of this approach on the development set are very different than with the other experiments. There are only values given for 1 and 5 answers returned because there are no more than 5 answers returned by the intersection method (the WordNet answers only coincide with at most 5 of the training method responses, frequently less). The intersection has a high precision rating (in part because of clusters where there were no labels in common between the two methods) but overall has low recall compared to the other variations. With just 1 answer returned, the recall is 28%, but the precision is 65%. If all the answers are returned the recall is 37% and the precision drops to 61.6%. The contrast is more clearly displayed in Figure 5.2 against the other experiments. Since WordNet on the Development Set has a maximum recall of 50%, the upper bound on the intersection recall is the same.

5.4 Test Set Results

The answers were evaluated against the human responses given for the Test Set. The same suite of experiments were run as with the Development Set. The Test Set did not have exactly the same ratio of Named and Regular Entities as the Development Set. Out of the 100 clusters, 34 of them were classified as Named Entity while the other 66 were Regular Entity clusters.

It is important to note that the data the computer uses to generate answers does not contain a possible answer for every cluster given the answer key generated by the humans. In fact, only 86 of the 100 clusters had a possible answer in the feature data, which means that a recall of 86% is the best the computer can hope for with this answer key.

5.4.1 Baseline Experiment

The same baseline experiment was completed as with the Development Set. On the Test Set, the baseline performed worse than it did in the Development Set testing. In Figure 5.3, the results of the baseline on the test set is shown compared with the other variations. With 1 possible label returned, the baseline found a correct label for 17 out of the 100 clusters. At 4 possible labels returned, the rate increased to 32 clusters correctly labeled. As with the Development Set test, the baseline does very poorly compared to the other methods, although the PRF set flags behind it after a few more labels returned.

5.4.2 Heuristic Experiment

The two sets of features introduced in Chapter 4 were also checked on the Test Set. The ADF Feature Set did better than the baseline, while overall the PRF feature set did about as well as the baseline. With a return of 1 possible label, the ADF feature set assigned correct labels to 18 clusters, and when returning 4 possible labels, 44 clusters correct. The PRF Set, on 1 possible label, started out with 22 clusters correct, but with 4 labels returned there was not as much improvement, with 34 cluster labels correct. In Figure 5.3, the

results for these sets is shown contrasted with others. The ADF set has similar performance here as it did on the Development Set, but the PRF set performs quite differently than it did previously. It starts off doing much better than all the other methods save intersection, and then abruptly sinks to the lowest position.

5.4.3 Basic Method Experiment

Next in the set of Test Set experiments, is the basic training method with no modifiers. The basic method was trained on all of the cluster data except for the clusters in the Test Set. Then the basic labeling method was run on all one hundred clusters in the Test Set. In Figure 5.3, the basic method is shown compared to the other experiments. With 1 possible label returned, the basic method produced a 17 correct labels, and with 4 returned answers there was 43 correct labels.

Like the baseline and the heuristic methods, the basic training method performs worse on the Test Set than it did on the Development Set. As before though, it does do a better job than the baseline method.

Further experiments were performed with the basic method on the Test Set. As in the Development Set experiments, to see if there is a difference between the Named Entities and the Regular Entities, the basic training was applied to the appropriate clusters. This time, there is 34 Named Entity clusters and 66 Regular Entity clusters. The results for these are not as different as they were with the development set. With 1 return, the Named Entity training produces 17.6% recall while the Regular Entities had 18.2%. Then with 4 possible labels returned, The Named Entities had 41.2% recall versus the Regular Entity result of 43.9%. The results of these tests are displayed in Figure 5.4, plotted against the results for the baseline under the same conditions. The baseline results on the two parts are also quite similar, although the Named Entities do better overall than the Regular Entities. The Baseline on the Named Entity part of the Test Set produce a recall of 17.6% with 1 possible label returned, and 32.4% with 4 possible labels returned. The Baseline on the Regular Entities performed a 16.7% on 1 answer returned, and a

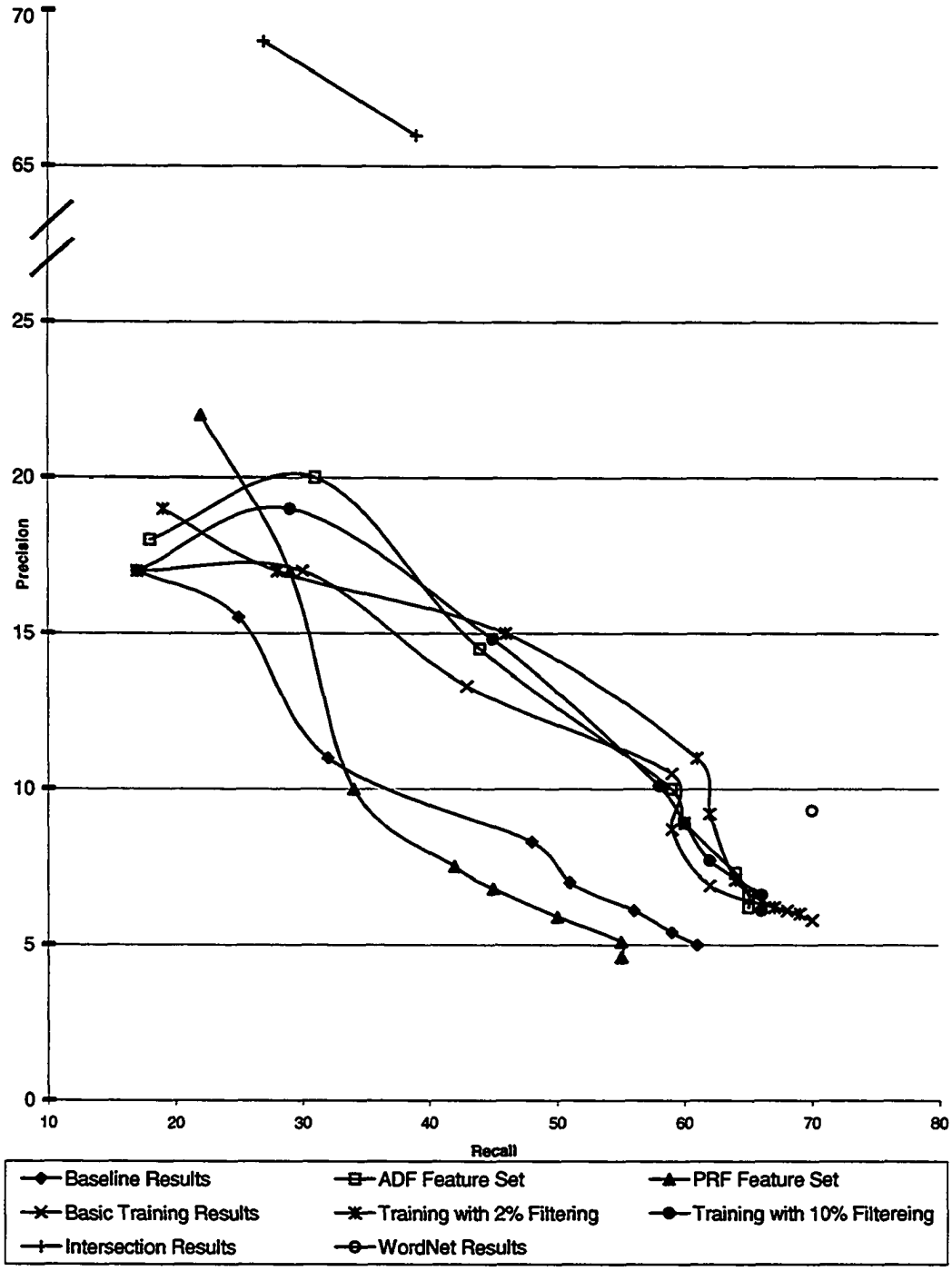

Figure 5.3: Results on the Test Set

31.8% with 4 answers returned.

In the graph, it is clear that the basic training method does quite a bit better than the baseline on both Regular and Named Entities, but there is not as strong evidence to say that Named Entities are easier to classify. The results from both types are far more intermixed than they were in the Development Set.

5.4.4 WordNet Experiment

As with the development set, the clusters in the Test Set are labeled with WordNet. Looking at these WordNet responses, they are considerably different than the results on the Development Set. The coverage from WordNet for the Named and the Regular Entities is about the same, unlike the previous values in the Development set. This is likely due to the difference in the numbers of clusters in each segment, and also because of the difference in how the answer keys were produced. The difference may also explain why the other methods did not show a substantial change in results between the Named and Regular Entity tests. WordNet produced a recall of 70% overall, with a recall of 70.6% on the Named Entity clusters, and a recall of 70% on the Regular Entity clusters. It would be expected that WordNet would have less coverage of the Named Entity clusters, due to the poor coverage of names in WordNet overall, but that is not the case here. WordNet has a much higher precision in the Test Set than it did in the Development Set as well, so either there was less ambiguity, or the human produced answer key simply had more of the terms that WordNet uses (something that is not too surprising an idea).

5.4.5 Frequency Filtering Experiment

Frequency Filtering is applied in a 2% and 10% magnitude, as during development these were the most advantageous. The results are shown against the other methods on the Test Set in Figure 5.3. The recall after the 2% filtering with 1 possible label returned was 19% (a small improvement over the Basic Training), and with 4 possible labels the recall went up to 46%. The 10% filtering did about as well, with a rate of 17% with one return, and 45% with

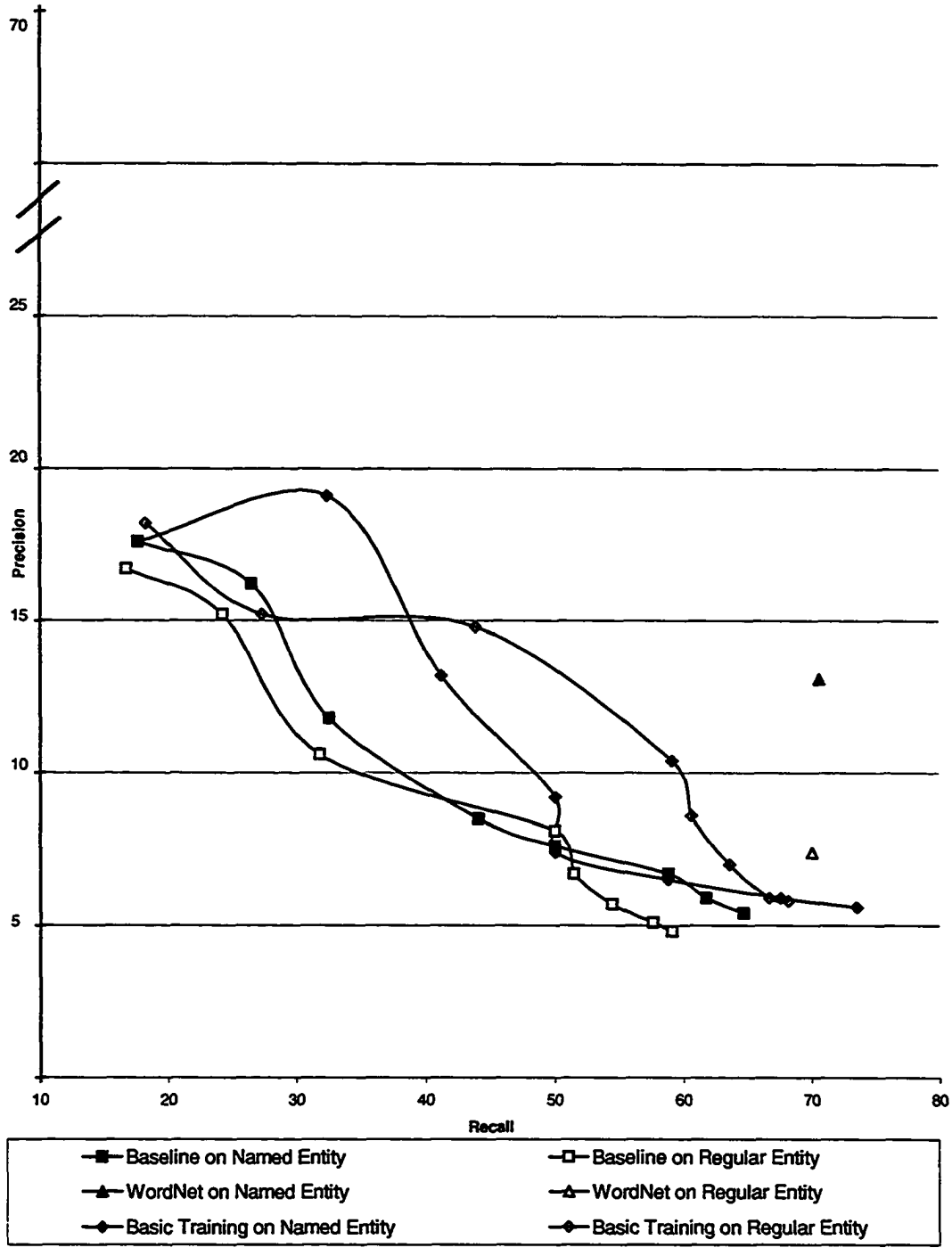

Figure 5.4: Results of Regular and Named parts of the Test Set

4 returns. In the Test Set though, the frequency filtering failed to overcome the ADF set until returning more than 4 possible labels.

5.4.6 Intersection Experiment

The intersection results for the test set are actually very close to the results from the Development Set. In this case, the first label given by intersection gave a 27% recall and a 69% precision, while the full answers gave a 39% recall and 66% precision. The results for the intersection method are shown in Figure 5.3 against the other methods. As with the Development Set, after 5 returns the intersection method can do no better, as it runs out of possible labels that occur in both the Basic Training Results and the WordNet Results.

It is not too surprising that the intersection does so well, given that 86 clusters are able to be labeled by the training method, and WordNet does so well on the Test Set.

5.5 Discussion of Experimental Results

The evaluation of the Test Set results is not very reliable as a way to determine the value of the computer generated answers. The human responses given in the study were often compound phrases, while the computer method mostly gives single word labels. As well, the human answers are not all correct, nor do they take into account the specificity of the answers. For example, while most names are easily labeled as referring to people by the human participants, the specific professions or roles that those people have in common are not so easily identified. There were frequent cases of the computer method identifying more specific labels than the human results did.

Since precision and recall do not take into account how specific the labels are, but the development set answer was purposefully chosen to be very specific, it is possible that the training method has better performance than can be seen with these regular measures.

It is also interesting to notice that WordNet did much better against the human responses than it did against the Development Set answer key. This

observation is not too surprising though, since WordNet is a human built tool. It makes sense that the language used by the participants in the study might have more in common with WordNet than the dependency data used by the learning algorithm.

The WordNet results also brings some strange properties of this particular Test Set to light. It is odd that WordNet would do just as well on the Named Entities as the Regular Entities in the Test Set, since WordNet in general does not concentrate on adding as many Named Entities.

As expected, the development set results far outperformed the test set results. Despite the differences in the magnitude, the training methods do much better than the baseline in both the Development Set and the Test Set.

Chapter 6

Conclusions and Future Work

The result of this work has been the creation of a new unsupervised learning algorithm to classify clusters of nouns with hypernym categories. The algorithm uses dependency information for the nouns in the clusters to determine appropriate labels for the cluster, and does not require the data to be annotated. As well as the basic algorithm, several variations were investigated, including an intersection method that combined the results of the algorithm with information obtained by WordNet. For the purposes of testing, a human study was also conducted to provide appropriate and unbiased evaluation of the answers given by the algorithm.

The unsupervised algorithm presented in this thesis has successfully labeled clusters of nouns. The method did substantially better than the baseline on the task. Some of the modifications, such as the frequency filtering, also improved the overall results. In particular, the intersection method produced a much better precision, but at a cost of reduced recall.

There are several possibilities for future work. One immediate expansion would be to add other types of concept groups to the mix, rather than just the noun terms. Adjectives and verbs are both present in Lexical Databases such as WordNet, so it would be interesting to see if a similar method would work as well to categorize them.

In order to better evaluate the performance of the algorithm, it would be valuable to see if it adds any increase to the performance of a Question Answering system.

Another valuable avenue of research lies in developing a standard for evaluating the results of cluster labeling (and other similar tasks of extracting semantic information). The common use of judges to determine correct responses does not easily take into account how "good" a label is, while more specific labels tend to be more useful. A new way to evaluate other than the standard precision, recall and f-measure could assist in evaluating these methods further.

Lastly, it would be valuable to discover if with the application of some automatic clustering method and a parser, this method could be applied to other languages. Automatic assistance could speed up the production of lexical databases for other languages.

Bibliography

[1] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL-99)*, pages 120–126, 1999.

[2] Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting of the Association for Computational Linguistics (ACL-85)*, pages 299–304, 1985.

[3] M. Fleischman and E. Hovy. Fine grained classification of named entities. In *Proceedings of 19th International Conference on Computational Linguistics (COLING)*, 2002.

[4] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of 14th International Conference on Computational Linguistics*, 1992.

[5] Nancy Ide and Jean Veronis. Knowledge extraction from machine-readable dictionaries: An evaluation. In *Proceedings of 3rd International EAMT Workshop on Machine Translation and the Lexicon (EAMT93)*, pages 19–34, 1994.

[6] Daniel Jurafsky and James H. Martin. *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, 2000.

[7] Dekang Lin. Latat: Language and text analysis tools. In *Proceedings of Human Language Technology Conference*, pages 222–227, 2001.

[8] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database, 1993.

[9] Patrick Pantel. *Clustering by Committee.* PhD thesis, Department of Computing Science, University of Alberta, 2003.

[10] Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *Proceedings of Human Language Technology / North American chapter of the Association for Computational Linguistics (HLT/NAACL-04)*, pages 321–328, 2004.

[11] Jason Rennie. Wordnet::querydata: a Perl module for accessing the WordNet database. http://www.ai.mit.edu/people/jrennie/WordNet, 2000.

[12] Stuart J. Russell and Peter Norvig. *Artificial Intelligence A Modern Approach.* Prentice-Hall, 1995.

Appendix A

Data example

N107 Prabowo Subianto, Wesley Clark, Prabowo

-N:person:N 1504

commander 199 GENERAL 1171 chief 134 -N:by:N 2 STATEMENT 2

-N:gen:N 73

visit 29 staff 8 move 5 comment 5 remark 5 aide 4 retirement 3

declaration 2 response 3 contention 2 replacement 2 status 3 Deputy 2

-N:with:N 76

meeting 46 talk 30 -N:from:N 8 request 6 commander 2

-N:of:N 88

command 17 resignation 20 dismissal 11 replacement 5 retirement 4

involvement 4 presence 7 endorsement 4 request 4 OFFICE 7 departure 3 support 2

-N:to:N 12

instruction 3 adviser 4 letter 5

-N:subj:N 62

commander 25 chief 11 HEAD 6 graduate 3 CHAIRMAN 6 minister 4

stranger 2 member 5

-N:conj:N 263

commander 47 GENERAL 77 chief 14 President 76 minister 33

secretary 7 white 4 meeting 3 parliament 2

-N:appo:N 419

commander 224 chief 50 President 56 leader 44 GENERAL 14

grandfather 5 he 11 minister 10 MUCH 2 them 3

-N:nn:N 4

all 4

N:nn:N 649

commander 167 Lt. 256 chief 91 army 108 Maj. 9 Brussels 3 blip 15

N:appo:N 598

commander 287 GENERAL 68 chief 45 military leader 4 HEAD 50

chief of staff 12 CHAIRMAN 84 leader 39 officer 7 military officer 2

N:conj:N 321

commander 68 chief of staff 19 GENERAL 54 chief 29 officer 22

Pentagon 4 HEAD 14 CHAIRMAN 15 official 29 secretary of state 5 others 20

governor 6 President 18 leader 12 lord 2 colleague 4

N:subj:N 4

ONE 4 N:on:N 11 WEDNESDAY 7 Tuesday 4

N:title:N 2334

Gen. 2334

end123

Appendix B

Feature List

-N:with:N	-N:of:N	-N:to:N
-N:subj:N	-N:conj:N	-N:subj-in:N
-N:appo:N	-N:nn:N	N:nn:N
N:appo:N	N:conj:N	N:subj:N
N:of:N	N:from:N	N:in:N
N:at:N	-N:over:N	-N:on:N
-N:by:N	-N:at:N	-N:in:N
-N:from:N	-N:for:N	N:to:N
N:for:N	N:gen:N	N:on:N
-N:below:N	-N:as:N	N:num:N
-N:against:N	-N:such as:N	-N:gen:N
N:as:N	-N:person:N	-N:outside:N
-N:among:N	-N:subj-on:N	-N:like:N
-N:via:N	-N:through:N	-N:subj-at:N
-N:into:N	-N:subj-for:N	-N:out:N
-N:between:N	-N:than:N	N:subj-in:N
N:than:N	N:after:N	N:with:N
N:around:N	N:between:N	N:by:N
N:over:N	N:subj-for:N	N:such as:N
N:UNDER:N	N:within:N	N:across:N
N:before:N	N:during:N	N:into:N
N:since:N	N:through:N	N:throughout:N
N:until:N	N:via:N	N:without:N
N:against:N	N:near:N	N:subj-on:N
N:among:N	N:appo-mod:N	N:about:N
N:along with:N	N:along:N	-N:about:N
-N:before:N	N:sub-about:N	N:regarding:N
-N:location:N	-N:inside:N	-N:subj-under:N
-N:after:N	N:aboard:N	-N:sub-behind:N
-N:sub-without:N	-N:prior to:N	-N:sub-like:N
-N:subj-among:N	-N:behind:N	-N:sub-of:N
-N:subj-as:N	-N:subj-about:N	-N:without:N

-N:throughout:N	-N:sub-to:N	-N:since:N
-N:during:N	-N:toward:N	-N:out of:N
-N:because of:N	-N:around:N	-N:per:N
-N:up:N	N:per:N	N:because of:N
N:instead of:N	N:subj-at:N	N:subj-to:N
N:like:N	N:subj-as:N	N:due to:N
N:behind:N	N:Despite:N	N:subj-like:N
N:prior to:N	N:all over:N	N:concerning:N
-N:subj-out of:N	N:person:N	N:inside:N
-N:near:N	-N:UNDER:N	-N:above:N
-N:off:N	N:next to:N	N:outside:N
N:beside:N	N:subj-among:N	-N:along with:N
N:off:N	-N:subj-from:N	-N:subj-down:N
N:out of:N	-N:subj-between:N	-N:towards:N
-N:within:N	N:beyond:N	N:according to:N
N:subj-of:N	N:subj-against:N	N:subj-with:N
-N:past:N	N:subj-from:N	N:title:N
-N:regarding:N	-N:according to:N	N:toward:N
N:guest:N	N:subj-under:N	-N:due to:N
N:down:N	-N:onto:N	N:onto:N
N:out:N	-N:concerning:N	-N:barring:N
-N:subj-against:N	-N:Despite:N	-N:subj-due to:N
-N:until:N	N:subj-between:N	-N:along:N
-N:down:N	N:past:N	-N:appo-mod:N
-N:across:N	N:subj-due to:N	-N:subj-by:N
-N:subj-around:N	-N:subj-off:N	-N:subj-up:N
N:up:N	N:above:N	N:towards:N
N:subj-behind:N	N:subj-out of:N	-N:beyond:N
N:self:N	-N:subj-over:N	-N:subj-than:N
-N:subj-below:N	-N:subj-above:N	-N:as against:N
N:subj-above:N	N:as of:N	N:thanks to:N
-N:amid:N	-N:all over:N	-N:next to:N
N:subj-off:N	N:beneath:N	-N:thanks to:N
-N:subj-near:N	-N:plus:N	N:amid:N
N:upon:N	N:subj-because of:N	-N:beside:N
N:location:N	N:plus:N	N:subj-without:N
-N:subj-inside:N	-N:subj-after:N	-N:subj-through:N
-N:upon:N	-N:subj-with:N	N:subj-around:N
N:subj-by:N	N:subj-near:N	-N:except:N
-N:aboard:N	-N:subj-within:N	-N:instead of:N
N:subj-within:N	-N:subj-before:N	N:subj-outside:N
N:subj-over:N	N:subj-before:N	N:subj-up:N
-N:except for:N	N:below:N	N:subj-down:N
N:subj-through:N	N:subj-below:N	-N:but:N

-N:subj-until:N	-N:guest:N	-N:content:N
-N:as of:N	N:subj-out:N	-N:alongside:N
-N:subj-out:N	N:except:N	-N:amongst:N
N:except for:N	N:pending:N	-N:subj-across:N
-N:subj-on to:N	-N:subj-during:N	-N:subj-aboard:N
-N:subj-into:N	-N:subj-all over:N	-N:subj-because of:N
-N:subj-beyond:N	N:but:N	N:subj-than:N
N:subj-until:N	-N:on to:N	N:round:N
N:subj-into:N	N:subj-during:N	-N:till:N

Appendix C

LaTaT Script

Here is a sample of the LaTaT script used to extract the feature information for the clusters (the script excerpt here is from the section to extract the development set info).

(make-tripledb cluster 200011) (link-file cluster cdep.hdr) (s cluster "N0 over 100, more than 200, more than 70") (s cluster "N1 plethora, slew, multitude") (s cluster "N10 Fahrenheit, Celsius, centigrade") (s cluster "N100 Novartis, BASF, Hoechst") (s cluster "N1000 Emmanuel Petit, Marcel Desailly, Bixente Lizarazu") (s cluster "N1001 BMG, EMI, Sony") (s cluster "N1002 television network, television station, network") (s cluster "N1003 MEDIA, reporter, journalist") (s cluster "N1004 Ind., Mich., Ill.") (s cluster "N1005 Los Alamos National Laboratory, LOS ALAMOS, Lawrence Livermore") (s cluster "N1006 Preakness Stakes, Preakness, Belmont Stakes") (s cluster "N1007 Olympia Snowe, Susan Collins, James Jeffords")

...

and so forth (there were one hundred clusters queried)

Appendix D

The Test Clusters

1) courage , determination , perseverance ,patience , dedication, tenacity, fortitude , resourcefulness ,selflessness, bravery, toughness , decisiveness, hard work ,resilience , foresight, natural ability , willpower , persistence ,wisdom , steadfastness, vision, resiliency , generosity ,diligence, LEADERSHIP, Statesmanship , spunk, self-discipline ,intelligence, forbearance, boldness , fearlessness , valor ,unselfishness, patriotism , self-restraint, willingness, gumption ,graciousness, chutzpah , moxie , open-mindedness, firmness, civic duty, stubbornness, self-sacrifice , will , tact, guile ,shrewdness, farsightedness, heroism , grit , sacrifice ,youthfulness, gallantry , thoroughness, penitence, combativeness ,LOVE, effort, resolve , stoicism, attentiveness, wile , depth ,humanitarianism, boosterism , trial and error, obstinacy ,follow-through, conviction, self-reliance, cowardice , volition ,hustle

2) Yasser Abed Rabbo , Nabil Shaath , Hassan Asfour ,Saeb Erekat, Oded Eran , Fehmi Agani, Martin McGuinness , Abu Mazen , Mahmoud Abbas, Erekat , Faisal Husseini , Hanan Ashrawi ,Long Yongtu, Yuri Maslyukov , Robert Aventajado, Anatoly Chubais ,Shi Guangsheng, Jaswant Singh, Eisuke Sakakibara , Yossi Beilin ,Haim Ramon, David Bar-Illan, Marc Hodler ,Sergei Yastrzhembsky , Xie Zhenhua , Yitzhak Molcho, Ariel Sharon ,Jibril Rajoub , Shaath, Peter Mandelson , Jacques Rogge , Richie Phillips, Boris Fyodorov, Mo Mowlam , Kevan Gosper, Ahmed Tibi ,colonel, Sadako Ogata, Dick Pound , Nur Misuari , Hans Blix ,Stanley Fischer , Danny Yatom , Dave Johnson, John Podesta , Wen Jiabao, Victor Malu, Abu Ala , Abdollah Nouri , Wang Zhongyu ,Wolfgang Schaeuble , Enrique Iglesias , fino, DAVID JONES , John Koskinen, Granik, Goldman Sachs Group , Berri, Bernard Kouchner

3) Feb. 1. , Feb 22 , March 7. , Feb 19, Feb 1 , March 7 , Feb 29 , Feb. 7., Nov. 6. , Feb 8, Jan 24 , Sept 12, June 7. , Feb 15 , Dec 12 , Nov 7 ,Oct. 2., Feb 27 , March 3., Nov 3, Sept. 5. , Jan 16 , Nov. 3. ,Aug. 5., June 6. , Nov. 2., Nov. 7. , July 5., Dec. 5. , Oct. 3. ,Nov 4, October 23, August 22 , Sept. 4. , February 10 , Aug. 9. ,October 18 , September 12, February 1 , Aug. 2.

4) New South Wales, Queensland , Western Australia, South Australia , Lower Saxony, Northern Territory, Tasmania , Victoria , Saxony, Hesse ,Oaxaca, Michoacan, Baja California , Sarawak , Yucatan, Sabah ,Sinaloa, Morelos, Jalisco, Tamaulipas , Zacatecas, Guanajuato ,Brandenburg, Chihuahua, Selangor , Guerrero , Bahia, Pahang, BAJA ,Durango, Nevis, Kedah , Nebraska, Australian state, Okinawa Prefecture, federal district, Hidalgo , PARA , Johor

5) Mario Elie , Jaren Jackson, Avery Johnson, Sean Elliott, Jerome Kersey ,Will Perdue, Terry Porter , Elie , Jeff Hornacek, tandem , Igor Larionov, Karim Abdul-Jabbar, Kerr , Hu Weidong, Tim Bogar ,swingman

6) master's degree , bachelor's degree, doctorate, law degree , Ph.D, degree,

B.A , MBA , laude, honorary degree, diploma ,B.S, knighthood , MASTER , course, major, graduation, Harvard Law School, GPA , undergraduate , bachelor, U.S. Naval Academy, degree program , expert, first degree, specialist, graduate student ,economics , Harvard College, education, researcher, training ,professor, physic, professorship , graduate, lecturer, cum

7) dugout , locker room, bench, sideline, clubhouse, LINE, fan ,history. , batter's box, uniform , dressing room, penalty box ,JERSEY , list. , Damien Woody, courtside, press box, sandlot ,huddle , orthopedist, backstop, lifer, midcourt, Jerome Bettis ,Rick Aguilera

8) mountain climbing, skateboarding , sledding ,surfing , gardening, riding, fishing, ski , sailing, jumping ,hunting , Whitewater, skier , Badminton, horse racing, table tennis

9) rightfield, leftfield , upper deck, second deck ,bleachers , grandstand, Pavilion , bullpen, fence, row , Stand ,gap, balcony , picnic area , chairlift, foul line

10) Josef Stalin , Idi Amin, Nicolae Ceausescu, Pol Pot, Stalin , Saddam Hussein, Franco, Hitler , Ceausescu, Mussolini, Pinochet ,Khrushchev, Khmer Rouge, Paul Coverdell , Kim Jong Il, Jim Henson ,Mobutu , banda HERBERT HOOVER , Emperor Akihito

11) Los Angeles Police Department, LAPD , New York Police Department, police department , NYPD, Scotland Yard, New York City Police Department ,New Jersey State Police, Royal Ulster Constabulary, State Police ,Street Crime Unit, child welfare agency , Internal Affairs ,PROVIDIAN, Rampart Division , New York City , sci, Environment Ministry , Bank of Israel , FTA, Public Security Bureau, taskforce ,General Services Administration, Legal Aid Society, Texas Supreme Court, Olympic Movement , Lotus Development Corp., Kenyan Government, Rideau , Joseph McCarthy, U.N. court, rampart ,District Council, Ray Evernham

12) Straits Times Index ,Weighted Price Index, Industrial Index, Straits Times Industrials Index , Composite Index, Hang Seng Index , Philippine Stock Exchange Index, Nikkei Stock Average, Tokyo Stock Price Index ,Second Board Index, Xetra DAX , CAC 40, Emas Index, NIKKEI , Hang Seng China-Affiliated Corporations Index, Hang Seng China Enterprises Index , Hang Seng, DAX , STOCKS, STI, futures price ,HSI , sub-index, Sensitive Index , fineness, sub-indices, orders ,scrip , soybean, Frankfurt, bullion, ordinary, ci , PSE , Brent ,STEADY , Mercury , ordinary shares, Glaxo Wellcome , Lagardere ,Barclays, Koji Ito

13) Shaquille O'Neal , O'Neal , David Robinson, Bryant , Duncan, Iverson, Garnett , Stackhouse , Jamie Feick, Abdur-Rahim , Fizer, Malone, Stoudamire , Mihm , Eric Snow ,Nailon , Wang Zhizhi, Gatling, Greg Ostertag , Mashburn , Geiger ,Gillom , Yao Ming, Wolters, Shawn Bradley , Collons , Allen Iverson, Arvydas Sabonis, Artest, Oliver Miller, Shaq , Feick ,Boselli, Gervin , Gadzuric, Lampkin, Clyde Drexler , Lisa Leslie ,Paul Pierce, Nate Newton, Larry Whigham , Jeffers, MacCulloch ,Dishman , mourning, Brian Grant , Divac , Michael Finley, Jason Arnott , Darren McCarty, Nick Van Exel, Antoine Walker , Shane Battier , Etcheverry, rook, Dan Gadzuric , Tyronn Lue, Twin Towers ,Weah , Detlef Schrempf, Michael Dickerson , Joe Sakic , Ruben Patterson, Matt Geiger, Teppo Numminen , Clifford Robinson, Chris Slade , Greg Anthony, Heinze, Mark Stepnoski , Jamal Mashburn ,Steve Heinze, Sun Wen, Newbury , Chris Anstey, Bobby Holik, Boban ,Peter Forsberg, Kelvin Cato , David Wesley , Scott Boras, Dana Stubblefield, Preki , Chris Porter, Bryon Russell , He/she, Cindy Parlow, Gary Trent , Anson Carter , putback, Sergei Fedorov ,Antoine Carr, Lieberthal, Rowley , Larry Smith , Tracy McGrady ,Ray Ferraro, Leetch , three-and-out, Bryce Drew , Kobe , Todd MacCulloch, Rodney Rogers , Dwight Yorke , Eldred, Jason Williams ,Joe Thornton

, Ankiel, Hollis , Robin Fraser , Tony Meola ,Reichel, Calvary , Yolanda Griffith, Tyrone Nesby , Lester Archambeau, Wally Szczerbiak, Patrik Elias , Ulf Kirsten , Vitaly Potapenko, Alen Boksic , Fontaine , Brian Lara, Falco , Neel ,Stuart MacGill, Angus Fraser , Shaun Pollock , Strindberg

14) Robin Yount , George Brett , Orlando Cepeda, Mike Schmidt, Carlton Fisk, Dave Winfield, Nolan Ryan , Paul Molitor , Eddie Murray, Bob Gibson , Tom Seaver , Tony Perez, Ryan , Barry Bonds, Steve Garvey ,Lee Smith , Steve Carlton , Wade Boggs, Dale Murphy, Tony Gwynn ,Jim Rice, Mel Stottlemyre , Brett , Cy Young, Yount, Fan Zhiyi ,Don Baylor, inductees , Telemaco , Tom Watson

15) notion ,concept , idea, definition , conception , theory, vision, legal principle , theory of evolution, basic concept , LINE ,categorization, nomenclature , theorem , IMAGE, lingua franca ,subgenre , teaching method, Word of God , conundrum, universality ,Style , TV dinner , intelligentsia, tort, embracing , doctrine ,premise

16) bikini , panty , underwear, thong , swimsuit ,underpants, pantyhose , bathing suit , bra, pajama , glove ,stocking, corset , undergarment , trunk, leather, Speedo ,knitwear, derriere , flannel , DKNY, boxer , spandex, plaid

17) east side , west side , Upper East Side, Upper West Side , south side , Park Avenue, Left Bank , north side , Central Park West ,Lower East Side , corner , Sunset Strip, edge, Columbus Circle ,skid row, Beacon Hill , campus , Sunset Boulevard, fringe , first floor , ground floor, block , Staten Island , floor, waterfront ,Manhattan , outskirt, Hainan Island, Long Island , promontory ,lakefront , North Shore , five acres, Route 1 , Grand Canal ,stretch

18) West Texas , South Texas, Central Texas, East Texas ,Southern California , Northern California, Warren County ,TENNESSEE , Nebraska, Panhandle , Hill Country , Catskill ,COLORADO , Bible Belt , Mississippi Delta, Mojave Desert , Bekaa ,Kono, prairie , Lancashire , glade, Negev, Fayette County ,Central Florida, highland , Top 20 , Essen, Williamson County ,Michoacan , Roanoke, Sylmar , working class

19) NTV , ORT , NHK ,ITV , TELEVISA , Univision, Telemundo , RAI, Channel 2, Channel 9 ,True North , Beta, CNA , Justice Party, Chris-Craft, Mana ,English language , Scripps, El Mundo

20) grassland , wetlands ,forest, woodland , pasture , marsh, marshland , virgin forest ,wetland, land , savanna , cropland, swamp , farmland , tropical rain forest, rain forest , paddy field , grazing land, rainforest ,coral reef , mangrove, meadow, Habitat , orchard, Wilderness ,desert , grove, field , chaparral, cultivated land, prairie ,parkland , wood, oases , bog, bio-diversity, Mojave Desert , old growth , sea lane, Greenbelt, RESERVE , tundra, hayfield , DMZ ,cattle ranch, Everglades National Park , catchment area , paddy ,land resource , Sawgrass, soil, biodiversity , Shennongjia ,permafrost, Biosphere, douglas fir , crevasse, Chattahoochee ,timberland , female body, coca crop , Zhangjiajie , Route 128 ,gulch , kelp , peat, East Kalimantan , greenway , fallow

21) dissension , discord, disharmony, disunity , divisiveness ,discontent, rift , dissent, jealousy, disarray , fissure , panic ,harmony , revolt, consciousness, bitterness , procrastinator

22) hibernation, stupor , slumber, coma , dormancy , trance, mode ,sleep , funk, obscurity , hiding , unconsciousness, remission ,torpor , orbit, shadow , cocoon , oblivion, mourning , crouch ,lull, huddle, denial , LETHARGY, dugout , nap , seclusion ,purgatory , shade, groundouts, GROOVE , reverie , UNDERGROUND ,ice age , past tense, neutral, rapture , Bloom , contortion ,repose , daze , burrow

23) Eighth Avenue , Lexington Avenue ,Seventh Avenue, Second Avenue , Third Avenue , Fifth Avenue ,First Avenue , Madison Avenue , Broadway,

Broad Street , Bowery ,Interstate 35, St., Route 128 , Fillmore, Boardwalk

24) impact ,effect , danger, risk , consequence , threat, possibility ,repercussion, likelihood, ramification , ravages , implication ,probability, hazard , recurrence, toxicity , pitfall , fragility ,peril, stigma , fallout, usefulness , potency , spillover ,unpredictability , inevitability , vagary, risk factor ,vulnerability , specter, relevance , prospect , disruption ,significance , wear and tear , burden, constitutionality, letdown ,cost overrun, aftershock , duplication , frailty, symptom, malady ,havoc, temptation , reverberation , dilution, attractiveness ,dependence , feasibility, encroachment, influence , enticement ,stimulation , omen , menace, avoidance, widening , absurdity ,leakage , gravity , deviation, casualty, PRESSURE , contamination ,mischief , vicissitude , benefit, abnormality , bad luck ,blockage, shortcoming , deprivation, health benefit, mutation ,exaggeration , vibration, impairment, shock wave , futility ,affront , pathology , aberration, hangover, blemish , blight ,aftermath , time bomb , sickness, discomfort, stench , vortex ,Diversification , pain , POTENTIAL, cruelty, complacency , onset ,allure , agony , doomsday, going away , yoke, life-threatening ,undesirable

25) detail , information, content, confidential information , data , specific, info, inside information , Material ,minutiae , gist , document, bookmark, tidbit , news article ,medical record , news , trade secret, elaboration , book , TOOLS ,RECORD , documentation, detritus, nitty-gritty , hearsay ,detailing, feedback , TV program , smut, statistic , rewriting ,knowledge, play-by-play, clarification, Webcast, folo , update ,file

26) tractor trailer , semitrailer, school bus, trailer ,boxcar , pickup, freight train , railroad car , rig, freight car ,moving van, cart, container , tanker, station wagon, plow ,Cellphones, DC-10, House International Relations Committee ,computer programming

27) privatisation, divestiture ,reorganization, divestment , dollarization, realignment ,SEPARATION, RESTRUCTURING , conversion, reshuffling, SPINOFF ,rationalization, modification , spin-off , evolution, reshuffle ,swap , renegotiation, handover , redesign , upgrade, breakup ,succession , splitting, filtering , revamp, reintroduction ,finding of fact , NTRA , changing, IFAD, long-awaited , cull

28) Fulton County , Travis County , Harris County, Jasper County ,Jackson County , Montgomery County, Williamson County , Suffolk County , Dutchess County, Jefferson County , Tarrant County, Essex County, Cobb County , Cook County, Monroe County, Johnson County ,Maricopa County , Broward County, Los Angeles County, Palm Beach County , Madison County, Erie County , Ventura County, Miami-Dade County, Warren County , King County , Jasper, Nassau County ,Gwinnett County , Orange County, Fayette County , DeKalb, San Diego County, Rockdale County, Fulton , Gwinnett, federal district ,Cobb , Morelos, Broward, Cherokee , Hague, ORISSA, Bronx

29) dependence , reliance, dependency, dependent, emphasis , PRESSURE ,bullishness, interest expense , drag

30) din , noise , roar ,sound , rumble, sonic boom, explosion , buzz, hum, noise level ,cacophony , echo, drone , clatter , flash, loud noise , thunder ,gunfire, whir

31) fresco , wall painting, Mural, mosaic ,stained-glass window, icon, decoration , canvas, Garden, scroll ,inscription , carving, mask , replica , float, settee ,seismograph , sconce, shower stall

32) cattle ranch, dairy farm ,plantation, estate, kibbutz , ranch, Vineyard, reservation, dacha ,collective farm, indian reservation , dairy, preserve, herd, King Ranch, Snake River , rancher

33) Junction, intersection ,juncture, crossroad , crossing, entrance, traffic circle , sea lane, confluence, exit , Shashi, foot, terminus, nunnery , toll plaza,

checkpoint

34) middle finger , ring finger , index finger ,pinkie , little finger, pinky, thumb , two fingers , finger ,forefinger , toe , tibia, knuckle , heel , glove, Palm, fingertip ,fingernail, paw, joint , clicker, ski , larynx, stem, upright ,Shane Dronett, upper, stump

35) ACCORD, AGREEMENT, pact ,cease-fire, ceasefire , treaty, truce , peace treaty ,Comprehensive Test Ban Treaty, MOU , CTBT , Armistice, Maastricht Treaty , START II, settlement, Kyoto Protocol, Anti-Ballistic Missile Treaty, Wye, Oslo Accords , Geneva convention, Lome Convention , FTA, memorandum of understanding, Uruguay Round ,compromise , peace, licensing agreement , North American Free Trade Agreement, NPT, customs union , protectorate, state of war ,general agreement, power-sharing, Stability Pact, ABM , Clean Air Act, Good Friday, signing , communique, Arms, STATEMENT ,disarmament, ruling, arms control , issue, tie, GOVERNMENT

36) sewer system , sewer, sewer line, water system, water line, sewage system, drainage system , toilet , water pipe, pipe, drain ,sidewalk, main, drainage , water company, control board, stadia

37) shamble, tatter , shambles, disarray, Jeopardy, limbo, high gear , flux, disrepair, ruin , natural state , midst, crisis situation, offing , state of war, infancy, doubt , thrall, synch ,paddy field , hock, quandary , peril

38) Lazaro Gonzalez ,great-uncle, Marisleysis Gonzalez, Lazaro, Marisleysis ,great-aunt, Uncle, relative , second cousin, Juan Miguel , Lawler ,Robert F. Kennedy Jr., Ethel Kennedy , Janice, Michael Skakel ,Maria Shriver , U.S. Immigration and Naturalization Service, Tin Oo, Elian , plantation owner, Kurth

39) No. 11 , No. 9 , No. 7 ,No. 6 , No. 8 , No. 12, No. 5 , No. 10, No. 4, No. 14 , No. 13 ,No. 3, No. 15 , No. 16 , No. 17, No. 2, No. 18 , No. 1, No. 5. ,No. 21 , No. 4., No. 22 , No. 19 , No. 8., No. 26 , No. 20 , No. 25, No. 7. , No. 3. , No.3, No. 23 , No. 24, No. 27, No.1 , No.2 ,No. 2., No.4 , par-3 , par-4, No. 1. , par-5, No. 99, number one ,No. , hole, par , Right-click

40) law, regulation, legislation ,constitution , provision , amendment, rule , statute, criminal law, antitrust law , code of conduct, bylaw, Bill of Rights ,stipulation , penal code, proposal, covenant , Basic Law, basic principle , price control , clause, international law, trade bill ,Measures, procedure , trade barrier , wording, code, protocol ,counsel law, standard, revision , zoning, First Amendment ,principle , safety standard, current system , tax policy , trade embargo, quota system , U.N. Charter , moratorium, legal system ,tax system , legal document, white paper , trade agreement ,economic system, precedent, fundamental principle ,attorney-client privilege, rule of law, mandate , sharia, norm ,Declaration of Independence, gun control, charter , red tape ,system , Standards , Terminology, plan of action , poison pill ,martial law, welfare state , abolition, oath, NAFTA , pilot program , framer, capital punishment, exemption , language ,etiquette , statute of limitations, advisories, Miranda ,memorandum , definition, minimum wage, Internal Security Act ,Vienna Convention, safety lock , Ten Commandments , immunity ,electric chair , ratification , Waiver, referendum , common law ,rationale, segregation , privilege, Handbook, Maxim ,implementation , tax, remedy , Islam, specification, quota ,mandatory

41) appeals court , court of appeals , U.S. Supreme Court, state supreme court , appellate court , high court, Supreme Court , court , lower court, Supreme Judicial Court , federal court , US Supreme Court, judge , SJC, Court of Appeal, Appellate Division , Texas Supreme Court, Circuit Court, trial court ,Constitutional Court , U.S. District Court, European Court of Human Rights , Court of Arbitration, Feerick, PANEL , District Court , US District Court, superior court , jury , Federal District Court, House of

Lords , Parades Commission , John Feerick ,European Court of Justice , U.S. Bankruptcy Court , Justice ,Court of Final Appeal , juvenile court, circuit, criminal court ,World Court , GC, Constitutional Council, U.S. Immigration and Naturalization Service, Rules Committee, International Court of Justice, General Services Administration, Congress Working Committee, ruling , American Psychological Association, insurance regulator , bench , Monetary Policy Committee, Patel ,International Amateur Athletic Federation, abortion law

42) sq.m , M. , cu.m, M , sq.km , ha, TEUs , km, kilometer, kwh , MU ,dwt, SQ , MW , meter, hectare , RAI , copy, yard , spindle , ft ,handset , about $20 billion , square foot, kinescope , megabits ,ATP Championship, more than 2 , tv set, telephone line

43) spoonful , Dollop , scoop, layer , mound, sprinkling, dab , Dash ,ladle, sprinkle , Drizzle , coating, gratin , about 1 cup ,Frosting, crystal

44) one tenth , One fifth , one-fifth, about one-fifth , about 44 percent , one-third, more than one third ,about one-tenth , One third, one-fourth, about 10 percent, about one-third, About 46 percent , thirds, one-tenth, about 26 percent ,about 15 percent , nearly one-third, about two-thirds, about 20 percent , about one quarter, nearly one third , About 50 percent ,about 28 percent, about 12 percent, about 36 percent, about 25 percent, about 9 percent , More than one-third, fraction, about one-sixth , About one third , about 43 percent, one-seventh, About 23 percent , about 35 percent, proportion, one-quarter, About 27 percent, almost one-third, more than one-fifth , about one-fourth ,bulk , about 17 percent, about 18 percent, tenth, one-eighth ,about three-quarters, about 13 percent , two-fifths, about 95 percent, nearly one-fifth, about 21 percent , About 42 percent ,one-sixth , about one-quarter, quarter, less than one-third , one quarter , roughly one-third, about 38 percent, nearly one-fourth ,fifths, four-fifths , over two-thirds, portion, about three-fourths, two-third , less than one-tenth, three-fifths ,three quarters, nine-tenths, almost three-quarters , one-half ,three-tenths, tenths , thousandth

45) kg , kilogram , ton, kilo ,five kilograms , Three Tons, one ton , two tons , six tons, one kilogram , two kilograms , Five tons, two pounds , metric ton ,five pounds, pound , kgs , a few pounds, kilowatt hour , bushel ,one pound, square meter , gram, liter, gallon , million yuan ,feet, one ounce , sachet , one hectare, bag , pfennig , four cents, sheepskin , four to five, gigahertz, acre , ounce, trillion dollars

46) LOW , highest level , HIGH, record level, closing price , offering price, RECORD , peak , rock bottom, loss, miles per gallon , high temperature, lowest , Ebb , dollar, mortality rate , two percent, second-highest, Three percent, about 18 percent , Nil, Celsius, lower , Fahrenheit, trough, mark , Est ,recovery

47) highest, largest , biggest, lowest, strongest ,smallest, steepest, cheapest , tallest, slowest, deepest, busiest ,fastest, richest, BEST, heaviest , lower than, Australian Bureau of Statistics, deadliest , weakest , Statistics Canada, hottest ,toughest , U.S. Commerce Department, same, shortest , less than one-tenth, about 36 percent , poorest, oldest, third one, another first , Central Election Commission, one sack , equity interest

48) turnabout, turnaround , reversal, about-face, shift , turn of events, u-turn, departure, improvement, comeback , change of direction , advance, leap, discovery, showing, ending , comedown ,revival, RETREAT, switch, thaw, burst , climb , turn, mercy killing , hush money, surrender, comeuppance , divergence , twist ,dry run, squandering , hot stock, nonevent , Zimbabwean dollar ,splurge

49) fulfilment, fulfillment , realization, attainment ,implementation, completion, pursuit , embrace , abrogation ,compliance

50) drug use, behavior , sex, alcohol abuse ,promiscuity , incest, crime, drug addiction , adultery, substance abuse , immorality, hooliganism, treachery , cannibalism ,pedophilia, child abuse, drinking, gunplay, smoking, drunkenness ,foul play, acts, sodomy, swearing, polygamy, blasphemy ,witchcraft, sexually transmitted disease, absenteeism , sexual assault, alcoholism, homosexuality, nastiness, incitement ,intoxication, contraception, suicide, cohabitation , masturbation ,debauchery, addiction , misogyny, obesity , masculinity ,consumerism, behaviour , harassment, extravagance, voyeurism ,insider trading , blood transfusion, gift-giving , disobedience ,self-destruction, USE , dieting, bodily function, deceit ,seduction , volunteerism, incontinence, lovemaking , autism ,attack , psychosis , sterility, materialism, sleaze , spitting ,kissing , racketeering , obscenity, alienation, banality ,hedonism, loneliness , evasion , taunting, toxic waste, alcohol ,mating, posing , cross dressing , chewing, whiteness

51) kind ,type , sort, variety , form , array, plethora, intricacy, panoply ,variant , hodgepodge , semblance, assortment, juxtaposition ,smorgasbord, compendium , profusion , confluence, checklist ,amalgam , gamut, ream , makings , welter, font, assemblage, trove ,crux , synthesis , New Style, broadening, subculture, cacophony ,mockery , epitome , abstraction, interplay , rigor, adjective ,contour , variation , permutation, patchwork, mishmash , new line ,preponderance , spectrum, trapping, version, cornucopia, geometry ,mastery , essence, constellation, potpourri , Nexus, guise ,subset , realm, concoction, lots, crescendo , tidal wave, life cycle , totality, thicket, myriad, purveyor , delicacy, evocation ,lifeblood, orgy, prism, freshness , taking, lieu, labyrinth ,sheaf, continuum , Dollop, copying, alchemy , full complement ,Crucible, granddaddy, flowering , detriment, series, melting pot ,surfeit, parameter, cutting edge , rendering, hotbed, hybrid ,annals, arbiter, offshoot , proliferation, specie, size ,initiation, description, reincarnation , simplest, antithesis ,Imitation, contraption, pattern , treasure trove, mouthful ,Convergence, reinterpretation, progeny, bane , patron saint, told ,conglomeration , larger, caprice, rehash

52) minister ,ambassador , foreign minister, envoy , secretary, secretary of state, deputy foreign minister, representative, Javier Solana ,undersecretary , minister of finance , finance minister, military attache , high commissioner, diplomat, William Daley ,vice-president , head of state, Kofi Annan , Robert Rubin ,delegation, governor general , Sheikh Hasina, Council of Ministers, William Cohen , Lawrence Summers , crown prince ,Wiranto , emir , Atal Bihari Vajpayee, Annan , cabinet member ,King Norodom Sihanouk, Saeb Erekat , Richard Butler, Consul ,Dennis Ross , premier , vice-premier, Bill Richardson ,president-elect , Albright, attache , Jean Chretien , Viktor Chernomyrdin, Yevgeny Primakov , Richard Holbrooke , emissary ,Holbrooke , councillor , Prime Minister, Ryutaro Hashimoto, Chuan Leekpai , Lionel Jospin, Nawaz Sharif , negotiator, mediator ,Vladimir Putin , Shimon Peres , FM, Aziz , Keizo Obuchi, John Howard, vice chancellor , tung , Goh Chok Tong, John Major ,al-Sabah , Solana, councilors , Romano Prodi , general staff ,Ivanov , Vajpayee , Wesley Clark, Moussa , Abubakar , President Kim Dae-jung, Li Peng , Jospin , Museveni, Weizman , Sergei Kiriyenko , Clerides, Daley , Qian , chairperson, Rafsanjani ,Hariri , counterpart, Queen Elizabeth II , commander in chief ,facilitator, councilor , Berger , Lebed, Anwar Ibrahim, Socialist Party , National Council, Zhu , national anthem , Jiang, Jiang Zemin , vice , national flag, sultan , International Committee ,troika , moderator , Cardoso, ex-president , Kim Jong Il ,stakeholder, executive council , Constitutional Court, European Parliament, Gross Domestic Product , wise man , Social Democratic Party, advisory board , Foreign Affairs , go-between, pm , CPC Central Committee , World Food Program, co-chaired, Ross ,Implemen-

tation Force, vice-chairman , Henry Kissinger, DEPT ,aegis

53) ITALY , France , AUSTRALIA, Spain , Sweden, Denmark ,England , Germany , Japan, Netherlands , United States, Norway ,Finland , Belgium , Argentina, BRAZIL , Russia, Switzerland ,Hungary , The Netherlands , Romania, Portugal, Bulgaria , Austria ,Poland , Lithuania , Slovakia, Croatia, Iceland , Czech Republic ,Ukraine , Greece , Latvia, Slovenia, Uruguay , Estonia, Great Britain , Chinese Taipei , Belarus, Trinidad and Tobago ,Luxembourg , Scotland, Ireland , Moldova, Liechtenstein, DPR Korea , Paraguay , Bolivia, Thailand, Malaysia , Malta, Wales ,Oceania , Ecuador, West Germany, Kazakstan , Czechoslovakia ,Scandinavia , San Marino , Indonesia, Republic of Ireland , Sri Lanka , Mallorca, West Indies, PAKISTAN , Peru, Maldives , Andorra ,Sicily, Yugoslavia , Slovak Republic , Philippines, Bavaria ,Vietnam , Belize, Olympiakos, Galatasaray , East Germany ,Greenland , Bayern Munich, Panathinaikos, Bhutan , Monaco ,Guadalajara, Mark Philippoussis, Tuscany , Juventus, sion ,Fiorentina , Queensland, Shanghai Shenhua , Boris Becker, Antigua , Real Madrid, Bosnia-Herzegovina , Parma, Manchester United ,Bayern , USA, Holy Land , Liverpool , AC Milan, home country ,Barcelona , Ajax, Boca , Bologna , Brittany, SIEMERINK, Todd Woodbridge, Kaiserslautern, U.S.A. , Dalian Wanda, Kafelnikov ,Tottenham, Aston Villa , New South Wales, Marseille, Sjeng Schalken, Sichuan, Valencia , Newfoundland , Southampton, Nazi Germany, Inter Milan , Ullrich, Rotterdam , Yevgeny Kafelnikov ,Waldner, South Pacific , Shandong , Henman, Jiangsu, Pat Rafter ,Fujian, Jim Courier , Leeds , Kuerten, Brisbane, the US , Turin ,Normandy , Monica Seles , Zhejiang, Steffi Graf, Newcastle ,Branson, Provence , Novotna , Trinidad, Liaoning, Glasgow, Minsk ,shuttlers , Kournikova , Kjus Bordeaux, Stuttgart, Martina Hingis ,Vero Beach , Agassi , Cali, Patrick Rafter , USSR, ROMA, Gama ,JELENA DOKIC , Andre Agassi, Napa Valley , New World ,Persia,roost, vs.

54) IAAF, FINA, International Amateur Athletic Federation, UEFA,UCI, FIBA, FIFA,UK Athletics , FIVB , International Amateur Athletics Federation, International Skating Union, FIA , Federation , Oca, IOC, International Cycling Union ,International Olympic Committee, DFB, Football Association , ITF ,USATF, EU Commission , U.S. Soccer Federation, International Tennis Federation , ITTF, ISU, FA , CAF, Asian Football Confederation , Olympic Council of Asia, English Football Association , F.A , FIS, CAA , PBSI , General Council, European Union , Koni , USGA, ACB , U.S. Golf Association, USA Basketball ,Association , Badminton Association of Indonesia, league , CFA ,Inter-Parliamentary Union, Academy of Motion Picture Arts , EBRD ,CONCACAF , National Executive Committee, fide, Council of Guardians , Swiss National Bank, UUP , CCPIT , State Peace and Development Council , England Club, King Mswati III, Aetna Inc. ,SPDC, Klestil , ulema , Guardian Council, AFC, Medical Society ,Hizbul Mujahideen, WBC , SEPA , Israeli Army, Banzer , WBA ,Cosatu, ATP Tour

55) liar , bigot , Traitor, hypocrite ,opportunist , racist, coward , perjurer , ENEMY, carpetbagger ,troublemaker , charlatan, denier , wealthy man, fascist, Evil ,prima donna , draft dodger, incompetent, interloper, manipulator ,scoundrel , monopolist , airhead, tyrant, deadbeat, misogynist ,SPY , crybaby , anti-Semite, obstructionist, philanderer, ladies' man, rube , cheat , hush money, lackey, Judas , hillbilly, outlaw ,sycophant , monster, making love, subversive, supplicant ,untouchable , accepting , Americanized, toady, toying

56) tone ,Style , attitude, mood , manner, atmosphere, demeanor, management style , personality, phrasing, temperament, Ambience, behaviour ,earnestness , sensibility, inflection, fluidity, diction, frame of mind , mien , timbre, good looks, ethos , gait, mentality, vibe ,ambiance, self-image , mind-set,

body language, mannerism, facial expression , persona, physicality , camera angle, coloration ,spirit , state of mind, worldview, world view, oratory , good humor, approach , pacing, behavior, asides, hairstyle, goings-on ,undertone , coolness, disposition, conceit, physique, complexion ,aura, professionalism , stillness, Romanticism, repartee , rhythm ,eroticism , VOICE , CADENCE, modus operandi, decor , nature ,artifice , dissonance, architectural style FASHION , spontaneity ,subtext, writing style, symmetry, coloring, Minimalism ,brightness , smugness, pretension, workmanship , sensuality, way ,intonation , milieu, color scheme, tic , pronunciation ,give-and-take , affectation , palette, gamesmanship , come-on ,connotation, proclivity , esthetic, wariness, Zeitgeist ,intensity , humor, drumbeat , patter, center of gravity, vibrato ,tenor , core group, chord , honesty, grammar, cleavage , gesture ,pastiche, accent , dialect, counterpoint, baritone , sonority ,vowel, glamour , silliness, populism, hue, overtone, atmospherics ,aplomb , veneer, articulation, decadence, character , theatrics ,countenance, Terminology , hyperbole, subtlety , originality ,styling, setting , SWAGGER , drawl, look, prose , self-confidence ,harmony , blandness , colour, detachment, locution , rhetoric ,tailoring , passion , narrative, scenery, play-calling ,underbelly, speech pattern , unison , vamp, confines , muttering ,lecturing, syncopation , glint , hush, A-list

57) plaudit ,accolade , kudos, adulation , Acclaim, praise, rave , scrutiny ,following, publicity , grade , ridicule, rating , scorn , airplay ,dividend payment

58) achievement, progress , success ,accomplishment , breakthrough , headway, outcome , attainment ,firsts, performance , key role , result, feat , good fortune ,fresh start, vindication , momentum, unfinished business convening , news event , bargaining chip, catharsis , martyrdom ,step, conundrum , vitality , hot issue, dawning , Holy Grail ,missing link, ascendancy , high-water mark, resumption, upper hand , comings , victory, refinement, outpouring , contribution ,hard sell , cash cow , nirvana, Second Coming , moneymaker, great year, co-leader , growth industry, no-no, brilliance, have-not ,Going, capstone

59) computer science , anthropology, sociology ,mechanical engineering, zoology, chemical engineering ,comparative literature , biology, mathematics, Science, economics ,political science, psychology, electrical engineering ,Engineering, linguistics , geology, social science, literature ,physic , biochemistry, chemistry, physics , liberal arts, math ,botany , astronomy, Microbiology, Criminology, astrophysics ,molecular biology , geography, agronomy, journalism, pharmacology ,physical education, civil engineering, social work, physiology ,natural science , theology, fine arts, Archaeology, architecture ,environmental science, neuroscience, information science ,medicine, English , nursing, civics, communication, earth science ,political economy, education, philosophy, accounting, archeology ,geophysics, ecology, Oceanography, elementary education , art ,paleontology, hydrology , fine art, folklore , humanity , anatomy ,FINANCE, language , ethic, theory, algebra , HEALTH, special education, bioethics , classic, law, meteorology , technology ,canon law, metaphysics , Public Affairs, urban planning ,statistic, veterinary medicine, particle physics

60) Hubbell ,McDougal, Cisneros, Steele , Hsia , Gotti, Neulander , Lee , Trie ,Bakaly, Tripp , Fassett, Susan McDougal, Hiett , Edwards ,Symington, LeFave , Julie Hiatt Steele, Volpe, Ickes , Espy ,Lacresha, Rideau , Schmalensee, Blackthorne , Mudge, Nichols ,SHARIF, Meskini , Demjanjuk, DeBartolo , Lyons , Lundy, Kimes ,Webster Hubbell , Mike Espy, Pirro , Mitnick , Kagalovsky, Nixon ,Piggie, Beckwith, Salinas , Maritz , Bruder, Haughey, Wen Ho Lee ,Barrionuevo, Grossberg , Lipstadt , client, Edwin Edwards , Maag ,Mohamed, Ray Lewis , Maria Hsia , Colburn, Combs, aide , Jarrell ,Qureshi , Biskind , Waters, king , Butts , Lawler, Farber , Chang ,Ratner, Sorensen , Trevi , Terry Nichols, Ford, John Kim , Claes ,Pugh , Gwen

, Ummi, accused , fink , Jobe, Hsieh , Bellush ,Milken, Theodore Kaczynski , Petersen, Traficant, Allchin ,Napolitano , Mellencamp, Bundy , Fayed, Gilliam, D-Ark.

61) pesticide , chemical , toxin, contaminant, PCBs , Material ,CARCINOGEN , insecticide , herbicide, DDT, fungicide , free radical, organic compound , allergen , gasses, impurity ,chlorofluorocarbons , radioactivity, MTBE , nitrate, malathion ,additive , food additive , PCB, residue , refrigerant ,preservative, coolant , tritium , uranium, compound , tremolite ,ester, Phosphorus , nutrient , aerosol, OD , DEET , solvent ,ingredient , electrical device , acid, alcohol , disinfectant ,talc, phentermine , liquid , ephedrine, Olestra , iodine , plant disease, FUEL , biomass , fluid, ammonia , sulfate , potassium

62) G-15 , SAARC , Rio Group, GCC , ASEAN , ECO, G-77 , SADC, APEC ,OIC , Group of 77 , NAM, EU , PECC , ECOWAS, COMESA , G8, D-8 ,Mercosur , CPLP , WEU, EAC , arf , OPEC Asem , G-8 , G-7, Andean Community , bloc , commonwealth, Andean Group , East African Cooperation, ci , ACP , KEDO, DLC , NDA , IGAD, ICC , ICA, Warsaw Pact, CGI , GEF , AIA, Rotary International , World Meteorological Organization, CE , subregion , Eximbank, UNIDO, South American Common Market, Olympic Movement , PTA , IFAD, Fla

63) HOME ,House , bungalow, property , farm , farmhouse, estate, mobile home , mansion, ranch , dwelling , vacation home , shack ,conference center , residence, ranch house , hut , hostel ,schoolhouse , shanty , campground, housing development , funeral home , storefront, beach house , dormitory , convent, Chateau ,brothel , office space, private property , barn , warehouse ,flats , country house , dorm, landfill , Fortress , homestead ,graveyard , shantytown , armory, greenhouse , family business ,Castle, silo , log cabin , cellar, service station , hideaway ,national monument, trailer park , mausoleum , amusement park, car ,hangar , store , manor , mailbox , lighthouse , campsite ,marina , courthouse, acre , golf course , army base, country club ,encampment , sports arena, convenience store , Holiday Inn ,bedside, Mount Carmel , dinner table , grotto, railway station ,windmill , house of worship, palace , trailer , Monticello ,parcel , lodging , orchard, airfield , islet , land, Staples Center , gymnasium , real estate, train station , sailboat , fort ,rink , compound , hive, theater , playhouse , racetrack ,headquarter , swimming pool , tent, quarry , ballroom, fairground ,vacationing , gym , ghost town, bakery , dealership, zip code ,multiplex , skyscraper , conference room, rose garden ,Rockefeller Center , monument, airport , grove , Safeco Field ,suburbia , gallery , Speedway, Auschwitz , Versailles , Alamo ,Eden , yearling , split up, two-story , Coop , Avalon, Victorian

64) opera , musical , cabaret, film , movie , Ragtime, Wild Party , operetta , musical comedy, Porgy , La Boheme , Annie Get Your Gun, Beauty Queen , Night Fever , Peony Pavilion, Sweeney Todd , puppet show , Les Miserables, Showboat , Evita , West Side Story, PLAY , Jekyll and Hyde , Marie Christine, scarlet pimpernel ,blackface , Via Dolorosa, Music Man , Riverdance , Phedre ,rainmaker , Yellow Submarine , Broadway, Lincoln Center Theater ,parade , misanthrope, Salome , Aladdin , Lion King, Hedwig , Act II , Smoke Signals, Neil Simon , Oedipus , Space Odyssey, Ideal Husband , Nightingale , Pimpernel, Werther , Bacharach , voiceover ,Royal Shakespeare Company , Clockwork Orange , Royal National Theater, Tin Pan Alley , blue angel , Eurythmics, high wire ,Pocahontas , high society, Gatsby , Little Mermaid , Jekyll

65) guitarist , arranger , vocalist, singer , pianist , saxophonist ,bassist , violinist , performer, clarinetist , bandleader , actor ,songwriter , drummer , actress, keyboardist , soloist , piano player, trumpeter , percussionist , guitar player, Rapper ,stuntman , MASTER, librettist , front man , musicologist,

pioneer ,frontman , long-distance runner, folklorist , beekeeper , Quincy Jones, John Mellencamp , Tony Bennett , Tito Puente, Top 20 ,string quartet

66) U.S. Open , Australian Open , French Open ,British Open , PGA Championship , United States Open, US Open ,Italian Open , PGA, Buick Open , Masters , Stockholm Open ,Wimbledon , U.S. Amateur , Grand Slam Cup, Canadian Open , Swiss Open , Paris Open, Buick Classic , LPGA Championship , Hong Kong Open, BellSouth Classic , Maurier Open , Chase Championships ,Acura Classic , Nissan Open , Loch Lomond, U.S. Senior Open ,PLAYERS Championship , Western Open, Buick Invitational , ATP Championship , Filderstadt, Dunhill Cup , GTE Byron Nelson Classic ,Lipton Championships , Tour Championship , London Marathon ,Maurier Classic , Roland Garros , New York City Marathon, Match Play Championship , ATP Tour World Championship, Skate America ,Maurier , Phoenix Open, Boston Marathon , British Opens , Ryder Cup, MasterCard Colonial , U.S. Opens , Solheim Cup, Safari Rally ,Eastbourne , Bay Hill, NEC Invitational , Giro , College World Series, OPEN , Indian Wells , Dinah Shore, U.S. , French Grand Prix , KITZBUEHEL, spelling bee , Doral , British Grand Prix ,major , Iditarod , NCAAs, Final Fours , Venice Biennale , Opens ,Ryder Cups , woman , SINGLE, Sendai , U.S. Women , Key Biscayne ,Forest Hills , Market Square Arena , Hilton Head, Indianapolis ,Longwood Cricket Club

67) National League , NL , American League, NFC , AFC , A.L., Al , league , Central League, major league , Pacific League , CWS, big league , AL. , Class AAA ,history. , Group A. , major league season, AHL , CAA , Midwest Regional, major , wildcard , West Regional, brave , International League , pre-Olympic, baseball , CAREER , East Regional ,Thomasville , Sydney Olympic , majors, Pacific Coast League ,Honus Wagner , two leagues, Arkansas Derby , Wenger , Ranger ,history , Wha , batting order

68) faith , religion , belief ,philosophy , tradition , ideology, religious belief, spirituality ,orthodoxy, Marxism , creed , mysticism, divinity, axiom ,creationism, theology , heritage , truism, superstition, atheism ,economic theory, pragmatism , liturgy , Deng Xiaoping Theory ,radicalism , humanism , adage, precept , environmentalism ,sanctity, piety , mythology , IDEAL, scientific theory ,demography , form of government, liberalism , American Dream ,multiculturalism, teaching , strong belief , religious movement ,pluralism , jurisprudence , myth, credo , evangelism ,fundamentalism, social system , iconography , scientific method ,Republicanism , qigong , manhood, tenet , Scripture , pacifism ,materialism , feminism , psychoanalysis, vested interest ,Puritanism , individualism, correctness , sect , ritual, virtue ,folklore , expediency, nicety , nationalism , cosmology, Chastity ,classicism , deity, modernism , mother tongue , individuality ,holy place , unionism , commandment, principle , Surrealism, root ,Spiritualism , fides , semantics, psychiatry , torah, medical history, God's Will , nobility , youth movement, missionary work ,custom , blackness, sacrament , sociology, papacy, dictum , Koran ,basic principle, political movement, gold standard , here and now ,ilk , occult , well-versed

69) Mazar-i-Sharif , Mazar-e-Sharif ,Jaffna, Jaffna Peninsula , GOMA, Anlong Veng, Kisangani , Makeni ,Grozny, Freetown , Bamiyan, Jalalabad, Kilinochchi , Bukavu ,Vavuniya, Uvira , Kenema, Atambua, Kindu , Kandahar , Kismayu ,GUDERMES , Juba , Huambo, Lubumbashi , Andulo , Herat, Argun ,Kailahun , Irbil, Gulu, Baidoa , Pailin, Bailundo , Camp Abubakar ,Adana, Arua, Srebrenica , Elephant Pass, Bo, Eastern Province

70) telephone number , phone number , Social Security number ,password, identification number , address, mailing address ,street address, name and address, Location , zip code , id ,signature, registration number , label, URL , blood type ,identifier, order form , return address , key, ticket stub , code ,marital status, wristband , false name , golden rule, bricks and mortar , tax

form , shoplifter, convenience food , torpedo boat ,Christmases, doorbell

71) varnish , enamel , sealer ,polyurethane , lacquer, stain, paint , sealant , epoxy , pigment ,adhesive , coating, wax , glue , grout, latex , coat of paint ,Mylar, primer, moisturizer , Grease, ink , remover , blood, coat ,putty, finish, mildew , glaze , plastic, endoscope , polyvinyl chloride, shaving cream, nail polish , nematode , powder, thinner ,chemical , lichen

72) stockholder , shareholder , creditor ,INVESTOR , policyholder , depositor, bondholder , employee ,minority shareholder, partner , holder , funders, whistleblowers ,London Club , sponsor, Lucio Tan , share , liquidator, ADM ,financial officer , capital investments, racketeer , Morgan Stanley Dean Witter & Co., personal loan , YPF , Viag, Amr ,Paris Club , small cap, Inhofe , parastatals , Coryatt

73) proof , indication , evidence, example , direct evidence, reminder ,inkling , result , indicative, circumstantial evidence, SIGN ,part, ONE , corroboration , exemplar, microcosm , warning signal ,intimation, foregone conclusion , harbinger , heartening ,wellspring , repudiation , modicum, good example , approximation ,test case, outgrowth , portent , perversion, byproduct , case in point , eye opener, instance , affirmation , predictor, sum total ,reference point , personification, refutation , validation, brunt ,confirmation , upshot , reaffirmation, signpost , clue, hint ,tradeoff , signal , choice of words, illustration , locus, bright side, latest , mirror image , corrective, watchword, paragon ,grist, progenitor , tip-off , distillation, elements, travesty ,oxymoron, invasion of privacy , fluke , kaleidoscope, stunner ,easiest , culmination, testament , disavowal , CASE, waste of time , exact opposite , vindication, dichotomy, motivator , piece of cake, apotheosis , ado , whopper, order of the day , bread and butter , dint, crazy quilt , dreg , chock, well-received , ebbing ,schematic

74) EPO , erythropoietin, human growth hormone, HGH ,steroid , testosterone, growth hormone , hormone , substance ,insulin , nandrolone , diuretic, melatonin , androstenedione ,Botox, adrenalin , Pfc , dyslexia, vitamin D , bodybuilding ,preventive medicine, positive, clotting , bidis, poll tax , folic acid , COUNTERFEIT

75) Gainer , decliner , advancer, loser ,advance , outspent, performer , FALLS

76) hiking , snorkeling ,boating, horseback riding , scuba diving , camping, skiing ,backpacking , rock climbing, bicycling , scuba , water sport ,jogging , boarding, swimming, adventure , bathing , diving ,racing , tennis , walking, yachting , back country , ballroom dancing, rowing , ice skating, ecotourism, cycling , snowboard ,golfing, flagstone , roller skating

77) working group ,committee , task force, steering committee , commission , PANEL ,council , taskforce , JMC, delegation , Kosovo Verification Mission, advisory board , work group , military commission ,Council of Economic Advisers, subcommittee , ethics panel, search committee , Joint Military Commission, interim committee , CSC ,staff, governing board, assembly , forum, PLO Executive Committee ,General Council, Sanctions Committee, Constitutional Assembly ,Business Council, Publicity Department, control board , House Government Reform Committee, IRB , directors-general , CMAG ,Philippine Senate, history department , Truth Commission, ethics committee, conferees , Industry Ministry, SFC , National Economic Council, Senate Finance Committee , Parades Commission , Senate Appropriations Committee, crew , House Commerce Committee ,Electoral Commission , inspectorate , Commission on Presidential Debates , Appropriations Committee, Independent National Electoral Commission , Federal Open Market Committee, core group , Ways and Means Committee, UNRWA , Chris Patten, National Election Committee ,Propaganda Department

78) ECA , ESCAP , IPU, UNCTAD , WEF, WIPO ,ILO , Itu , FAO, Inter-Parliamentary Union , OECD , WTO, IEA, CCF ,UNEP, chamber of commerce , BOI , OAPEC, PATA , TDB, NSC, HKTA ,Bi , IAEA, TDC , AmCham , Cass, Cosatu , IDB , Dar, SPF, UPC ,ICTR, ATA , Thai Foreign Ministry , IPC, International Ski Federation , ai , Rainbow/Push Coalition, AA, Bea, State Environmental Protection Administration , Olympic Committee, South Pacific Forum , Strauss-Kahn , IAI, Central Organ, Hearst Corporation , wi

79) Dasa , Aerospatiale Matra, AEROSPATIALE ,British Aerospace , Lagardere , Daimler-Benz, Vivendi ,Thomson-CSF , Renault SA, Bae , Elf Aquitaine , Canal Plus, AXA ,ALLIANZ , Thomson CSF, casa , Vickers , Arianespace, ALCATEL ,Granada , Lufthansa, Heike Drechsler , Erik Zabel, Michael Stich ,Xoom.com , David Prinosil , NEC Corp., Heinz-Harald Frentzen ,MITSUI , New York-New Jersey MetroStars, Heidelberg , Bundestag ,Hubert Vedrine, South African Communist Party , Linares

80) indiscretion , misbehavior , misdeed, transgression , dalliance ,escapade, wrongdoing , peccadillo, involvement, affair , scandal ,infidelity, sin , infraction, conduct, dealings , liaison ,mistake , tryst , offence , lapse, excess , encounter , failing ,plot , lying , episode, misadventure , antic , wrong ,indecisiveness , Exploit , shindig, fling , hijinks ,prevarication, MP3s , shortsightedness

81) middle class ,wealthy , affluent, rich , elite , upper class, Kamel , savvy ,Meeks, married couple , Nature Conservancy

82) most valuable player , MVP , rookie, Defensive Player , NL MVP, NL Rookie ,rebounders , Doak Walker , African Cup, SURPRISE, Emmy ,vote-getter, Washington Mystics , Gold Gloves, triple-double ,runner-up , COACHES , series., baserunner

83) third period ,second period , first period, final period , second overtime ,second half, first half , first overtime , overtime, regulation time , THIRD QUARTER , second quarter, last two minutes , sixth minute , second minute, seventh minute , quarter, first quarter ,OT , first 10 minutes , third minute, period, extra time , extra inning, first 20 minutes , fifth minute , half, first minute, last five minutes , first two quarters, first 30 minutes , second and third quarters, first century , halfcourt, span, afternoon ,regulation , first five minutes, stretch, first nine , Pacific Division, nine minutes , waning

84) whisky, whiskey , cognac ,wine , tequila , red wine, Champagne , scotch, vodka, bourbon ,sparkling wine , Calvados, gin , Dom Perignon, alcohol, booze ,Madeira , bubbly, Bordeaux wine , mustard, bottle, moonshine ,spirit , sugar, spice , hot stock , cheese, pastrami , kimchi ,garlic lubricating oil , Curacao , pleasantry, aperitif , agave

85) worshipper , worshiper , parishioner, congregation ,congregants , church member, churchgoer , faithful, non-Muslims ,Scott McNealy , Vo Van Kiet , Sydney Organizing Committee ,trailblazers , organist , regencies, Caritas , North Carolinian ,Coloradan, day laborer , backbencher , aggregation, subatomic particle , Aurelio , Rick Neuheisel, World Council of Churches ,Scarpetta , Rodney Eyles

86) Agricultural Bank of China , China Construction Bank ,Industrial and Commercial Bank of China , Construction Bank of China, Bank of Communications , Bank of China , ICBC, State Development Bank , Commercial Bank of China, CCB , Eximbank , SDB ,Shanghai Branch , BOC , Macao Government, industrial , People's Bank of China, Hang Seng Bank , China Southern Airlines, PICC, Air China , Bank of Montreal, Beijing Municipal Government, Xinhua ,Shanghai Stock Exchange, Hong Kong Monetary Authority, CITIC ,SANWA BANK, executive agency , Xinhua News Agency , China Eastern Airlines, Helen Clark , UNICOM , SUMITOMO BANK, Criminal Investigation Department , PBOC

87) Swan Lake , Giselle ,sleeping beauty, Nutcracker , Don Quixote , Romeo, Ballet ,Macbeth , Turandot, repertory , masterpiece , Streetcar Named Desire, classic , True West , Iceman Cometh, Cinderella , Carmen ,Aida, Don Giovanni , Cleopatra , Othello, Otello , Shakespeare ,Peter Pan, KING LEAR , Faust , La Traviata, pas de deux ,Anastasia , Electra, Tristan , Amadeus , waltz, Dido , Tosca ,Rigoletto, Isolde , Juliet , firebird, Anna Karenina , Great Expectations , Dracula, chorus line , Spartacus , Falstaff, Julius Caesar , Hamlet , Snow White, Chekhov , takedown , title role ,Ben-Hur , Lulu , Antony, Taming , mini , baroque

88) Parcells ,Belichick , Holmgren, Fassel , Gailey , Reeves, Pitino , Monson ,Donnan, Tuberville , Schottenheimer , Carroll, Tomey , Romar ,Groh, Jauron , Troussier , Hackett, Wannstedt , DiCicco, Scioscia ,Jackson , Muckler , Bill Parcells, Guthridge, Krzyzewski , Kearin ,Tom Coughlin , Shanahan , Popovich, Passarella , Schmid , Chan Gailey, Polian , Penders , Rick Neuheisel, Riley , Harrick , Bill Belichick, Ainge , Larry Robinson , Al Groh, Billick , Dave Campo ,Rich Brooks, Ponciano, Vermeil , Franchione, Sampson , Rick Pitino , Van Gundy, Breda Kolff , Braswell , Chris Palmer, Solich ,Hartsburg , Kevin Gilbride, Bora Milutinovic , JOHN ROBINSON ,Gullikson, Stoneman, Spurrier , Zagallo, Brian Billick , Tim Floyd , Scotty Bowman, Vince Tobin , Mike Holmgren , Dungy, Phil Jackson , Ftorek, Calipari, Tobin , Ganassi , Piniella Jim Harrick ,Bowden , Marty Schottenheimer, Jim Fassel , Neuheisel , June Jones, Calhoun, Hiddink , La Russa, Milbury , Vogts , Norv Turner ,Ditka , Bavasi, Cowher, Sherrill , Wilkens , Skiles, Tubbs , Dick Jauron, Zambrano, Dunleavy , Dennis Erickson , Martz, Gruden, Pete Carroll , Lou Holtz, Ohlmeyer , Slocum , Babcock, Zampese, Danny Ainge , Reinsdorf, Tommy Bowden , Keady , Cowens, Lamont, Mike Ditka , Pat Burns, Steve Spurrier , Blazevic , Dan Reeves, Aliotti ,Hodgson , valentine, Cremins , Graziano , George Karl, Barry Switzer , Kurt Rambis , Dean Smith, Jacquet , Dom Capers, Jimmy Johnson, Ken Hitchcock , McCarver , Gullit, Switzer , Don Nelson ,John Calipari, Silas , Sather , Izzo, DAVIE , Joe Paterno, Dennis Green, Steve Mariucci , Tony Dungy , Evernham, McReynolds, Dave Wannstedt , Bochy, Mike Shanahan , Dick Vermeil , Whalen, Weisman ,Pat Riley , Ewbank, Jeff Van Gundy , Esrey , Holzman, Mike Krzyzewski , Auerbach , Beamer, Rudy Tomjanovich , Bill Walsh ,Shula, Walter Zenga , Issel , Whitsitt, Shurmur , Eskew, Paul Silas, Gorman , Pat Gillick , Bill Cowher, Bouton , Harr, Dimon ,Bela Karolyi , Ernie Zampese , Ray Rhodes, Glenn Hoffman, Jerry Sloan , Joe Gibbs, Bobby Bowden , Butch Carter , John Thompson ,Bubb , Auriemma , Brezhnev, Ruud Gullit , Larry Brown, Esiason ,Tarses , Robert Kraft , Eckersley, Begala , Jim O'Brien, Fiorina ,Mike Dunleavy , Spina , Kilborn, Brisby , Beathard , Tom Osborne ,Chuck Daly , Gossage , Johnny Oates, Bettman, Milutinovic, Muti ,Angelos , Childress , P.J Carlesimo, Hartwig, Don Shula, Johan Cruyff, Sykes , WEISEL , Jim Calhoun, Kevin Keegan, Ferrell , Alex Ferguson, Harry Sinden , Tuiasosopo, Donna Shalala, Cowan , Bobby Grier , Lehrer, Bear Bryant, Tarpley , Caray, McGowan , Figgis ,Kevin Kennedy, Bruce Bochy, Seiji Ozawa, Braugher, Liman , Hough ,Phil Garner, tiller, Doug Melvin , Roy Williams

89) mother ,FATHER , wife, husband, parent , brother, sister , mother-in-law ,stepfather, Aunt, grandmother , fiancee, grandparent , Uncle ,grandfather, exwife, stepmother, sister-in-law, boyfriend , MOM ,ex-husband, spouse, dad, father-in-law, brother-in-law , fiance ,Girlfriend, sibling, in-law , ex-girlfriend, Barton , widow ,half-brother, daughter-in-law , stepson , Princess Diana, best friend, son-in-law , Matthew, nanny , Daddy , Juan Miguel Gonzalez ,great-uncle , DiMaggio , Laura, little brother , mistress , Susan ,sweetheart , extended family , O'Connor, Diana , sitter , Tommy ,Sarah , roommate , Shannon, Linda , mama , hank, Edward , kin ,William, Hillary , Eric , Abraham,

patriarch , great grandfather ,Michael, Grandma , Robert , Kennedys, Ben , Clarke , playmate ,Anna , Arthur , Lazaro Gonzalez, Nancy , aristocrat , Andy ,tipper , Richard , Princess, Moses , Brandon , Steele, Patrick ,Nicholas , McDougal, duchess , Ramon , Joe DiMaggio, Steven , Amy ,Payne, Abiola , Woodward , Karen, Frankel , Matt , youngest ,Silverman , Rudy , paramedic, Anthony , Billy , Jacob, Dylan ,John F. Kennedy Jr. , namesake, Maggie , Charles , Jake, Assad ,Jackie , Linda Tripp, Kate , fireman , Tripp, Hussein , Paul ,accuser, great grandmother , Chung , Andrew, Hassan , earl, Peter ,Byrd , Joseph , Bulger, Ali , Leo , Gandhi, married man, Robbins ,Nathan, housekeeper , deceased , Christopher, mentor, Aaron ,Jennifer, MARY , Shepard , Kaufman, Queen Elizabeth, companion ,Gotti, hooker , Daniel , widower, matriarch , Havel, Justin, Fred ,Raymond , Adam, Lisa , Weaver , lover, Mother Teresa , Maria ,matron, Jonathan , Pedro , ballplayer, ancestor, Panchen Lama ,buddy, Spielberg , Serena , Volpe, Albert, hairdresser , Benjamin ,Rockefeller , Hemingway , chauffeur, family man , Kenny , Malcolm ,JFK , Leon , Tracy, little sister, Gabriel , Jesse, mourner ,Ralph , Leonard, Lucy , monarch, Bryan, Jessica , Gilbert , Lin ,Joan , Lyons , ALAN, Allison, accomplice , heir, xu , Payne Stewart , Joey, Jane , Doc, Stephen, ROCKWELL , Big Brother ,Marilyn Monroe, Rudolph , SEAN, Springsteen, TRUMAN , John F. Kennedy , pontiff, Josh , Kerry, hostess, Jefferson , Ted Turner ,Weston, George Washington, Alice , STUART, mao , real estate agent , Sonny, Thomas Jefferson, Chan , Ruth, Tyler , Jeb Bush ,Bernstein, Huang , pal , Wu, emperor , Sammy , alter ego, Sinatra ,God , Prince, entourage, stein , godfather, hippie , shah ,Salinas, Elizabeth , abuser, Uncle Sam, scion , Louis , Emily ,Jose , Lee , pastor, Hillary Rodham Clinton , Cher , furrow ,Oprah Winfrey , Murdoch , Charlie Brown, John Glenn , Posada ,Monroe, Michael Jackson , Lennon, lance, guardian , oldest, Woody ,Raul Salinas , idol , Muhammad Ali, living together, Mickey ,Mohammed, lawyer , Donald, attorney, first name , best man ,sidekick, Frank , Shaq, policeman, comrade , Hitler , Spice Girls ,co-defendant , Bashar, Santa Claus, George Bush , forebear ,identical twin, Willy, Alvin , Einstein, caddy , flipper , Mother Nature, Fang, chestnut , Barney, Negro , Newt , pre, mum , I. ,eldest, brood, buster , fairy

90) ballistic missile , missile ,warhead, surface-to-air missile , ICBM , nuclear warhead ,launcher, air-to-air missile, intercontinental ballistic missile ,S-300, chemical warfare , scud , MIG-29s, guided missile , THAAD ,W-88, cluster bomb , reverser , rocket engine, air-to-surface missile, Trident , avionics, stinger , tomahawk , torpedo poison gas, chemical agent , materiel, test-firing , silo , SHIP, air defense, Zvezda , proton, land mine , munition , Paranthan

91) Carter Holt Harvey , Fletcher Paper , Fletcher Energy, Fletcher Building, Auckland Airport , Lion Nathan, Brierley Investments ,Fletcher Forests , Brierley Investment, Contact Energy , Air New Zealand, telecom, Telstra , amp , Sky TV, warehouse , Fletcher ,New Zealand, tower , colonial , Nathan, Sanford , Qantas , Ansett

92) taxiway , tarmac , landing strip, runway , lane , gate, dock ,apron , ramp, deck , control tower , Diaoyu Islands, hangar ,Garden State Parkway , layer cake, airfield

93) vessel , SHIP ,tanker, freighter , cargo vessel , cargo ship, fishing boat, boat ,oil tanker, merchant ship , trawler , container ship, passenger ship , barge , supply ship, tugboat , patrol boat, gunboat, New Carissa , pleasure boat , ferryboat, torpedo boat, hovercraft ,landing craft, cruise ship , sailing ship , small boat, USS Enterprise , speedboat , schooner, icebreaker, Concordes ,helicopter, sailing vessel , oil rig , Exxon Valdez, ocean liner ,airship , drilling rig, car , luxury liner , glider, hydrofoil ,dhow , vehicle, Snow Dragon , Kursk , submersible, slave ship ,steamboat , tractor trailer, FERRY , riverboats, frogmen, train ,pontoon , lifeboat, dump truck , riverboat, steamer, Grand

Princess , Hindenburg , DC-9, ice floe , houseboat, Hunley ,commuter train , MD-80 , Sayonara, whaler , U.S. Coast Guard ,main line, transporter , Mayflower , cable car, life jacket ,boater , liner, pallet , Zeppelin , Yorktown, Britannia, flight deck , Queen Mary, Voyager , bogie , il

94) Akili Smith, Tim Couch , Daunte Culpepper, Donovan McNabb , Cade McNown, Brock Huard, Shaun King , Peyton Manning , Michael Bishop, Charlie Batch ,Doug Pederson , Ricky Williams, Ryan Leaf , Tee Martin , Brian Griese, Kerry Collins , Quincy Carter , Chad Pennington, Rob Johnson , Gus Frerotte , Keith Brooking, Dave Brown , Tom Brady ,Joe Hamilton, Doug Johnson , Corey Chavous, couch, Jim Druckenmiller , Drew Henson , Rich Gannon, CURTIS ENIS, Andre Wadsworth , Tony Banks, Damien Woody , Manny Malhotra, Jarious Jackson, Keith Smith , Danny Kanell , Jeff Blake, Chris Canty, L.J Shelton , Mike Van Raaphorst, Chris Claiborne, Ebenezer Ekuban ,John Avery, Chris Miller , Pat Burrell , Greg Ellis, Marcus Outzen , Solomon Page , John Tait, Drew Brees, Charles Woodson ,Thomas Jones, J.D Drew , Robert Edwards , Bob Hallen, Ephraim Salaam , Raef LaFrentz , Ortege Jenkins, Drew Bennett, Flozell Adams , Cory Paus, Wane McGarity , Todd Collins, Mike Bibby, Jason Fabini , Patrik Stefan , Tony Simmons, Michael Myers, Daylami ,Ryan McCann

95) fifth , sixth , fourth, seventh, eighth , third ,ninth , second , bottom of the inning, first, inning. , one-two ,over .500 , first point , third one, fifth and sixth , overall ,first to third, fastest , seventh and eighth, first turn, GAA ,No. 4 , furlong, No. 7 , runner-up, sixth and seventh, No. 3, No. 1 , No. 2, EDT , won three , Mika Hakkinen, No. 5, second-fastest ,TOP, AL West , run, homestretch, three points, NL West , Jose Offerman, Orlando Palmeiro , first six, won five, Eleventh, Jesper Parnevik, Shannon Stewart, PLACE, hundredth

96) equanimity, good humor, humility, aplomb , wit, grace, sincerity, precision ,modesty, maturity , restraint, sensitivity, humor, bravado , zeal ,poise, innocence , savvy, assurance , SWAGGER, intensity ,nobility, courtesy , optimism, regularity , relish, clarity ,alacrity, TEMPER , iron fist, flair , composure

97) car horn ,Siren, horn, whistle , bell, buzzer, noise , cowbell, alarm, fire alarm , church bell , chime, beep , roar , wailing, gong , salute ,trumpet, drum , bagpipe, mayday, dirge , firework

98) Mitch ,GEORGES , Hurricane Georges, hurricane , Hurricane Mitch , Debby ,Hurricane Bret , Hurricane Floyd , Floyd, Hurricane Bonnie , Babs ,Zeb, Bret , Bonnie, Hurricane Dennis, Dennis, Irene , George ,Alberto , Lenny, Bertha, Celia , Cesar , virus, Arlene , crest ,Charley, Hugo, Isis , scare, Jose , Fran , Vicki, Danielle ,Andrew , Flight 111, Tszyu , Kurd

99) apartment building ,condominium , apartment, condo , town house , townhouse, building ,apartment house, brownstone, townhouses , office building, studio apartment, carriage house , OFFICE , row house, duplex, walk-up ,tenement, high-rise , loft , hotel room, rooming house, resort hotel, penthouse, chalet , housing , tower suite , Miller Park ,co-op, houseboat , studio , manse, summer house , Belmont Learning Center, shop , Grand Hotel , space, skyboxes , Ebbets Field, Elysee Palace, flat , block , dacha, vacant lot , Oakland Coliseum , beachfront, Alcoholics Anonymous

100) Ruben Patterson ,Rashard Lewis , Vladimir Stepania, Vernon Maxwell , Tyronn Lue ,Jelani McCoy, Brent Barry , Horace Grant, Greg Foster, Devean George , Aaron Williams , Billy Owens, Raul Ibanez , Steve Park ,Russ Ortiz, Barry , Keith Smith,

Appendix E

Aggregate Human Results

The following responses are the result of the 8 participants labeling the 100 test clusters. The numbers following each label are how many participants gave that response.

1 attitude 1; manner 1; feeling 1; state of mind 1; virtue 1; personality trait 2; human trait 1; human quality 1; noun 1; character
trait 2; personal characteristic 1; personality trait 1

2 person 1; people 1; name 1; leader 1; leader in their field 1; men 1; no answer 2

3 date 4; calendar date 2; random date 1; day of the year 1

4 region 1; state 2; province 2; location 2; state 1; country 1; geographical area 1; territory 1; district 1; geographical location 1;
place 1; area 1; place in asia 1

5 person 1; athlete 2; sportsman 1; people 1; name 1; basketball player 2; people name 1; no answer 1

6 degree 2; post-secondary education 1; title 1; academic level 1; university 1; college 1; education level 1; skill level 1;
achievement 1; university degree 1; post secondary 1

7 location 1; sport venue 1; sports term 2; game place 1; sports 2; sports people 1; sports thing 1; found in a stadium 1

8 sport 4; pastime 1; physical activity 1; leisure activity 1; activity 2; hobby 2; outdoor activity 2; recreational pursuit 1

9 location 1; places in stadium 1; baseball term 1; locations in a baseball stadium 1; stadium place 1; viewing area 1; seating
area 1; stadium area 1; football 1

10 tyrant 1; politician 1; strongman 1; authoritarian 1; dictator 5; head of state 1; world leader 1; political figure 1;
puppet master 1; political leader 1; bad people 1; leader 1

11 institution 1; department 1; crime fighter 1; law organization 1; legal organization 1; public protector 1; law enforcement 1;
investigator 1; investigation agency 1; law enforcement agency 1; judicial system 1

12 stock index 1; index 1; stock market 2; trading 1; financial investment term 1; market index 1; stock exchange 4

13 person 1; sportsman 1; athlete 2; sports players 1; professional athlete 2; famous people 1; name 1; tall 1; no answer 3

14 person 1; baseball player 3; name 1; hall of fame 1; sportsmen 1; no answer 1

15 concept 1; notion 1; thought 1; theory 1; idea 1; logical construct 1; revolutionary idea 1; abstract concept 1; law 1; no answer 1

16 clothing 3; female clothing 1; revealing garment 1; foundation 1; underwear 2; clothes 2; undergarment 1; lower body clothing 1

17 location 3; address 1; urban location 1; New York state 1; New York city 2; place 1; city location 1; area 2; city place 1; New York 1

18 region 1; county 1; place 1; USA place 1; locations in the US 2; geographical area 1; country music 1; southern U.S.A 1; no answer 1

19 TV station 3; television station 1; public television 1; media 1; communication channel 1; television channel 1

20 environment 2; habitat 1; location 1; ecologic zone 1; natural environment 1; land type 1; area of cultivation 1; ecological niche 1;
nature 1

21 disagreement 1; state 1; state of affairs 1; emotion 2; disorganization 1; problem 1; emotional state 1; trait 1; negative group
dynamic 1; negative feeling 1

22 state 1; state of body 1; sleep 1; non-active state 1; state of mind 1; state of being 1; lack of alertness 1; seperate 1; no answer 1

23 location 1; urban location 1; street 5; address 1; road 5; route 1

24 feeling 1; emotion 1; sensation 1; cause and effect 1; consequences 2; negative result 1; unhealthy 1; no answer 2

25 information 4; media 1; information source 1; record 1; information store 1; data 1; journalism 1; communication 1; unit of
information 1; hospital 1

26 vehicle 5; mode of transportation 1; method of moving 1; transportation 1; cargo transportation 1; transportation method 12

27 action 1; process 1; change 1; streamline 1; corporate sales 1; corporate terms 1; reorganization 1; method of change 1; business
organization 1; economic euphism 1; cabinet 1

28 county 4; places in the USA 2; place 1;area 1; district 1; location 1; police department 1; geographical area 1; geographical
district 1; U.S. counties 1; vacation place 1

29 attribute 1; character trait 1; interpersonal interaction 1; dependancy 1; being a commander 1; no answer 2

30 noise 2; sound 4; level of noise 1; decibel 1; noise pollution 1; audio 1; can hear these 1

31 piece of art 1; interior art 1; wall treatment 1; decoration 1; interior design 1; building decorations 1; art 2; image 1; tiles 1;
visual display 1; art 1

32 land area 1; land property 1; farming 2; rural area 1; non-urban locations 1; property type 1; slavery 1; confined ecological
space 1; farming; farm type 1; growing thing 1

33 junction 1; traffic point 1; highway 2; road work 1; infrastructure 1; road terms 1; transportation feature 1; pathway 1; stopping
point 1; no answer 1

34 finger 1; body part 2; hand and arm 1; extremity 1; appendage 1; hand part 1; hand 2; foot 1; no answer 2

35 agreement 5; pact 2; treaty 4; political agreement 1; contract 1; war 1

36 sewage utility 1; underground utility 1; pipeline 1; infrastructure 1; water infrastructure 1; utility 2; plumbing 1; plumbing
infrastructure 1; water waste 1; water 1

37 state 1; state of disorder 1; problem 1; negative state 1; state of being 1; trouble 2; bad situation 1; danger 1; no answer 2

38 relative 1; people 1; cuban crisis 1; 2nd degree relative 1; second degree relative; Kennedy 1; no answer 3

39 number 3; ordinal number 2; position 1; video game ranking 1; golf 2; game ranking 1; direction 1; option 1; natural number 1

40 law 2; regulation 1; government 1; legal terminology 2; legal term 1; legal document 1; human rights 1; standard 1; regulation 1;
government term 1; legislative term 1; lawyer 1
41 court 4; legal system 1; legal organization 1; justice 1; law 1; court 1; governing body 1; death penalty 1
42 unit of measure 2; measurement 3; unit 2; approximate measurement 1; linear measurement 1; area measurement 1
43 quantity 1; baking term 1; cooking 1; measurement 1; amount 2; approximate baking measurement 1; cake baking 1; portion 1; cooking
term 3
44 fraction 5; less than one 1; part of a whole 1; part of 1; measurement 2; approximate fraction 1
45 unit of measure 2; units of weight 1; measurement 2; weight 2; bag size 1; measure 2; amount 1; mass 1; unit 1
46 level 1; rate 2; ups and downs 1; amount 1; benchmark 1; price 1; cost 1; measurement 1; no answer 1
47 superlative 3; magnitude 1; degree 1; statistic 1; section 1; extreme 1
48 change of direction 2; turn around 1; change 1; metaphoric movement 1; stock market 1; stock exchange 1; change 1; result 1; business 1; no answer 1
49 completion 1; realization 1; happiness 2; outcome 1; achievement 1; positive result 1; reward 1; goal 1; no answer 1
50 abuse 1; vice 1; sin 2; impropriety 1; negative connotation 1; taboo subject 1; negative human trait 1; controversial issue 1; moral
issue 1; behavior 1; taboo 1; action 1; personal trait 1; immorality 1; personality defect 1; mob 1; no answer 1
51 mix 1; random group 1; food 1; no answer 4
52 politician 1; diplomat 2; representative 2; government official 1; governor 1; political appointment 1; leader 1;
international government agent 1; international government agency 1; UN concern 1; government 1
53 country 3; team 1; contestant 1; place 1; location 1; international location 1; country with tennis stars 1; political area 1; county 1; no answer 1
54 sport organization 2; acronym 1; international sports association 1; organization 4; association 1; sports association 1
55 pejorative personal term 1; epithet 1; negative character trait 2; putdown 1; description 1; character type 1; traitor trait 1;
interaction 1; bad people 1; sinner 1
56 personal attitude 1; quality 1; exterior characteristic 1; expression 1; appearance 1; state of mind 1; no answer 2
57 praise 2; review 2; assesment 1; evaluation 1; recognition 1; positive reinforcement 1; music industry 1
58 achievement 1; accomplishment 1; investment term 1; performance 1; key economic indicators 1; winning 1; best 1; climbing the
ladder 1; no answer 2
59 science 1; university program 1; studies 1; learning 1; subjects of formal study 1; occupation 1; discipline 1; field 1 training 1; post
secondary study 1; area of study 3; university subject 1; university 1
60 person 1; surname 1; last name 1; people 2; spy 1; name 1; criminal 1; famous people 1; no answer 1
61 chemical 4; contaminant 1; chemistry 1; pollutant 1; toxin 1; chemical compound 1; biological compound 1; dangerous chemical 1;
harmful material 1

62 bloc of countries 1; political organization 1; acronym 1; organization 2; international group 1; international committee 1;
grouping 1; economic group 1; international association 1; group of nations 1

63 building 3; inhabitable building 1; places to live 1; location 1; property 1; gathering place 1; place 1; structure 1; lodging 1

64 musical 1; play 1; theatre 1; performance term 1; performance 1; performing art 1; off-broadway 1; entertainment 1; drama 1; show 1

65 musician 4; people 1; performer 2; entertainer 1; performing artist 1; music 1

66 tournament 1; golf tournament 2; sports event 2; competition 2; contest 1; sports tournament 2; sports competition 1; race 1

67 league 1; baseball 2; baseball term 1; sporting organization 1; sports team 1; sports group 1; baseball; hall of fame 1; no answer 1

68 personal belief 1; thought 1; world view 1; religion 1; idea 1; social term 1; belief 1; philosophy 1; belief system 2; controversial
topic 1

69 location 3; town 1; place 2; mass murder 1; genocide 1; ethnic cleansing 1; muslim 1; no answer 2

70 personal information 1; personal identification 1; label 1; index 1; identifier 1; identification code 1; identification 1;
identity 1; no answer 1

71 coating substance 1; liquid repair item 1; substance 1; coating 1; finish 1; covering 1; chemical covering 1; painting supply 1; no answer 1

72 shareholder 1; banking 2; finance 1; corporate term 1; stakeholder 1; corporate malfece 1; stock exchange 1; no answer 1

73 evidence 2; theory 1; experiment 1; conclusion 1; demonstration 1; indication 1; courtroom 1; no answer 3

74 hormone 2; biological substance 1; growth hormone 1; body 1; health term 1; bio chemical 1; drug testing 1; illegal sports drug 1;
prescription 1; no answer 1

75 stock attribute 1; stocks 3; character type 1; stock market 1; game 1; no answer 2

76 pastime 1; recreation 1; sport 2; activity 2; outdoor hobby 1; exercise 1; athletic activity 1; extreme sport 1; physical activity 1;
outdoor sport 1

77 committee 4; political organization 1; group 1; team 1; board 1; government sub-group 1; government group 2

78 organization 5; acronym 3; group 1; no answer 1

79 company 1; airline 1; organization 1; aerospace corporation 1; international corporation 1; no answer 3

80 transgression 1; sin 1; descriptions of public figures 1; negative human action 1; behavior 2; immoral behavior 1; adultery 1;
crime 1; no answer 1

81 division of society 1; social class 1; descriptions of wealthy people 1; social status 3; class 1; class structure 1;
classification 1; status 1; well off 1

82 category of player 1; athlete 1; professional sport 1; sports term 1; position 1; winners 1; sports designation 1; sports
achievement 1; baseball 1; hall of fame 1

83 time period of a sport game 1; sports game 1; timekeeping 1; span of time 1; part of game 1; game segment 1; sports event
division 1; basketball 1; hockey 1; game time 1

84 alcoholic drink 2; alcoholic beverage 1; food 1; drink 1; beverage 1; loosening agent 1; festive occasion 1; festive meal 1; ethanol 1;

dinner party 1

85 church member 1; people 1; noun 1; religious participant 1; creationist 1; christian 1; religion 1; no answer 1

86 bank 3; organization 1; financial institution 2; chinese stock exchange 1 ; chinese economic group 1;place to put your money 1

87 theatrical piece 1; stage production title 1; ballet 2; play 1; theatre 2; entertainment term 1; performance 1; classics 1; stage

performance 1; screen performance 1; broadway 1

88 surname 1; people 2; name 1; athlete 1; coach 1; hockey player 2; no answer 1

89 family relation 1; relative 4; people 2; relation 1; famous people 1; people; dead relative 1; role 1; name 1; president 1; royalty 1

90 missile 2; weapon 3; weapons of mass destruction 1; modern warfare 1; warfare term 1; nuclear weapon 1; warfare 1; military weaponry 1; ammunition 1

91 company 3; new zealand 1; new zealand corporation 1; no answer 3

92 airport component 1; airport 2; airport term 1; pavement 1; paved area 1; airstrip 1; part of airport 1; aircraft 1; no answer 1

93 vessel 1; boat 2; ship 3; ocean vessel 1; transportation 2; vehicle 2; shipping 1; water vessel 1; cargo; part of ship 1; ship part 1;

sea travel 1

94 person 1; people 2; name 1; sportsmen 1; football 1; no answer 3

95 inning 1; sports 1; baseball 2; horseracing 1; sports term 1; standing 1; ranking 1; placement 1; no answer 1

96 positive attitude 1; character trait 2; human quality 1; congeniality 1; personal characteristic 1; positive character trait 1;

goodness 1; child raising 1

97 noise-making device 1; noise 3; noise-maker 1; warning sound 1; signal 1; sound 1; loud noise 1

98 hurricane 4; disaster 1; storm 2; noun 1; quick thing 1

99 inhabitable building 1; residence 1; living quarters 1; building 2; urban location 1; dwelling 1; place to get drunk 1; location 1;

living place 1; place to live 1

100 person 1; people 2; name 1; male name 1; male people 1; men 1; no answer 2

Appendix F

Computer Cluster Labels

Here are listed the first ten labels given by the computer (basic training used) on the test set. The number following each label is the score it obtained. Labels in bold are labels that match the answer key.

1. policy 130.0 time 116.1 quality 111.5 heart 111.0 will 106.7 principle 99.5 strength 96.0 strategy 94.5 commitment 94.5 work 92.9

2.negotiator 104.7 minister 97.7 chief 90.8 official 90.8 deputy 90.4 **leader 90.4** aide 74.0 head 64.0 saeb erekat 63.4 member 62.5

3. primary 52.5 **date 43.0** deadline 34.4 week 32.1 victory 24.9 ballot 24.9 death 24.9 game 24.9 race 24.9 contest 22.4

4. **state 158.7** queensland 63.4 south australia 63.4 australian capital territory 63.4 governor 56.4 coast 49.9 premier 49.9 capital 49.9 minister 49.9 town 49.5

5. mario elie 63.4 david robinson 63.4 jaren jackson 63.4 each 55.1 player 46.2 @@ points 37.6 steve kerr 32.6 avery johnson 32.6 addition 32.6 he 32.6

6. study 150.7 **college 132.1** student 132.0 program 130.5 **university 126.4 degree 121.9** education 116.6 course 107.0 two 98.3 experience 97.2

7. team 82.1 field 77.2 crowd 74.9 red 72.4 edge 66.8 offense 66.6 locker room 63.4 pole 63.4 guy 61.9 everybody 61.6

8. **activity 103.2** event 90.0 **sport 82.1** shooting 74.9 hiking 63.4 swimming 63.4 tennis 63.4 skiing 63.4 golf 63.4 horseback riding 63.4

9. seat 103.9 tower 63.4 store 57.8 concern 53.2 race 52.9 he 44.7 wall 42.8 garden 39.4 field 39.4 one 38.7

10. regime 98.1 **leader 78.9 dictator 52.2** support 39.8 army 32.6 government 32.6 hitler 30.7 milosevic 30.7 president 30.7 dictatorship 29.3

11. officer 63.8 office 63.4 investigation 51.0 unit 49.9 chief 49.9 official 46.2 veteran 43.0 review 43.0 policy 43.0 city 34.7

12. @.@@ percent 55.1 currency 30.7 @@@.@@ points 22.4 @@.@@ points 22.4 @.@@ points 22.4 @.@ percent 22.4 future 22.4 futures contract 22.4 rose 22.4 @.@@ point 22.4

13. player 132.3 @@ points 127.0 star 108.7 pick 94.4 guard 93.2 each 85.8 scorer 78.0 center 77.6 forward 76.1 all 70.6

14. pitcher 93.2 george brett 63.4 robin yount 63.4 carlton fisk 63.4 roger clemens 63.4 willie mays 63.4 all 55.1 each 55.1 candidate 46.2 player 41.9

15. word 116.9 design 111.3 **idea 102.1** strength 99.0 **concept 98.1** view 97.1 one 95.4 president 94.2 character 93.4 influence 90.9

16. **underwear 85.8** bikini 77.7 top 76.3 suit 67.5 sock 63.4 skirt 63.4 sweater 63.4 pant 63.4 heel 63.4 shoe 63.4

17. **area 105.9** building 94.5 apartment 88.4 district 88.4 home 82.2 center 81.5 street 78.2 site 71.2 line 70.5 feet 67.7

18. area 128.5 state 90.9 penn state 63.4 michigan 63.4 nevada 63.4 wisconsin 63.4 baylor 63.4 oklahoma state 63.4 new york 63.4 game 58.9

19. television network 63.8 network 63.8 **television station 43.6** director 43.0 newspaper 32.6 tv network 29.8 station 29.3 interview 26.5 television 22.4 news 22.4

20. area 200.3 forest 145.5 resource 138.0 protection 133.2 plant 129.8 soil 121.9 land 117.6 park 113.4 population 108.2 road 108.2

21. fear 84.0 confusion 63.4 controversy 63.4 tension 63.4 unrest 63.4 protest 63.4 violence 63.4 **problem 46.2** danger 36.1 reason 33.1

22. blip 62.3 form 53.2 change 45.0 one 43.0 all 39.2 flight 34.9 you 32.9 color 32.6 streak 32.6 blush 32.6

23. corner 73.7 **street 69.2** broadway 63.4 entrance 59.4 building 57.5 avenue 48.1 side 43.0 block 38.6 store 35.2 madison avenue 32.6

24. one 152.5 problem 130.5 health 125.8 economy 122.3 market 121.2 factor 119.9 risk 118.2 war 116.8 disease 113.5 impact 113.4

25. page 159.4 **information 152.8** archive 151.6 report 150.9 research 149.3 one 129.4 service 127.6 number 124.2 president 121.3 site 119.2

26. truck 92.1 car 81.6 door 80.6 tank 65.9 **vehicle 63.4** trailer 63.4 bus 63.4 automobile 63.4 room 63.4 driver 49.9

27. plan 112.0 issue 85.8 management 84.9 **change 75.4** program 73.7 effort 70.6 idea 70.6 consolidation 63.4 expansion 63.4 sale 63.4

28. city 85.8 area 59.8 attorney 55.1 one 54.7 district attorney 49.5 sheriff 43.0 coroner 43.0 resident 43.0 chairman 43.0 north 37.9

29. problem 80.7 accountability 63.4 result 42.7 economy 41.1 community 34.8 growth 34.8 change 33.0 intensity 32.6 impulse 32.6 weakening 32.6

30. **sound 113.4 noise 106.5** light 74.9 firework 70.6 echo 68.7 all 64.8 scream 63.4 hiss 63.4 smoke 63.4 band 63.4

31. painting 125.3 statue 84.0 piece 80.8 flower 70.6 feet 68.7 sculpture 63.4 fresco 63.4 relic 63.4 carving 63.4 picture 63.4

32. land 101.5 road 89.1 island 77.7 place 75.4 vineyard 74.9 center 72.8 field 65.9 property 65.2 farm 63.4 estate 63.4

33. road 84.4 north 74.0 opening 53.2 site 53.2 side 53.1 **highway 52.2** door 44.7 street 44.2 east 43.0 west 43.0

34. **finger 106.5 hand 105.1** back 90.7 bone 82.2 injury 78.9 blip 77.8 thumb 75.4 eye 68.9 nose 67.5 wrist 63.4

35. issue 173.4 negotiation 164.8 **agreement 144.4** document 141.8 obligation 137.6 framework 136.9 principle 134.6 talk 132.1 protocol 126.3 commitment 123.5

36. pipe 97.3 **water 92.5** building 84.0 drain 78.9 network 75.7 construction 73.7 road 67.5 wall 65.7 sewer 63.4 faucet 63.4

37. risk 51.3 mayhem 37.9 one 34.2 that 32.6 more 32.6 variety 32.6 distress 30.7 chaos 30.7 scandal 30.7 kid 30.7

38. family 91.3 cousin 78.9 elian 67.8 great-uncle 63.4 uncle 63.4 brother 62.3 custody 53.4 marisleysis gonzalez 46.2 marisleysis 46.2 others 44.1

39. wood 147.6 hole 128.4 pick 113.5 no. 1 113.4 team 103.0 seed 100.1 ranking 96.7 duval 90.9 bogey 86.5 player 85.8

40. **standard 204.9** reform 196.1 state 188.4 issue 187.3 idea 155.4 debate 134.0 concept 130.5 review 128.2 campaign 125.0 decision 123.2

41. judge 194.5 decision 172.4 ruling 170.9 verdict 153.7 panel 139.5 **court 128.6** case 116.3 hearing 109.5 appeal 108.1 chief justice 104.3

42. area 123.5 land 60.3 building 53.8 zone 51.3 @@.@ percent 48.1 note 46.2 two 45.0 space 43.3 forest 37.9 one 36.7

43. photograph 63.4 sauce 60.3 each 57.2 strawberry 39.8 mixture 37.9 water 37.9 green 37.9 top 36.7 surface 33.3 slice 32.6

44. @@ percent 96.2 number 94.1 @ percent 86.1 more 85.8 half 77.2 $@@@,@@@ 75.4 share 70.5 output 65.5 population 64.9 woman 64.9

45. pound 105.9 cocaine 71.8 two 63.4 pant 63.4 **weight 60.4** more 55.4 symbol 55.1 number 51.3 floor 51.2 capacity 50.2

46. @@ percent 119.6 **level 119.2** number 109.2 average 104.3 **rate 99.7 price 89.4** dow 83.8 china 80.6 credit 78.9 @,@@@ 75.6

47. one 63.3 @.@ percent 57.8 @@.@ percent 57.7 @ percent 57.7 last 55.3 increase 54.2 history 53.0 period 46.2 team 43.8 @.@@ percent 42.2

48. **result 135.3** improvement 93.2 surprise 89.7 something 80.2 success 78.9 rise 75.4 market 71.3 effort 70.5 victory 69.8 step 66.8

49. resolution 78.9 process 78.9 task 65.5 agreement 64.2 implementation 63.4 promotion 63.4 progress 53.2 strategy 42.7 administration 42.7 development 39.8

50. problem 166.2 act 139.6 violence 125.2 drug 118.4 one 118.2 issue 117.4 practice 113.8 treatment 112.0 rate 112.0 incident 106.4

51. one 177.6 program 134.9 character 125.9 world 124.8 book 121.2 number 121.0 music 119.2 kind 115.1 image 108.5 ones 105.9

52. member 188.2 president 187.8 meeting 186.9 **government 179.5** visit 174.2 general 168.2 official 167.4 chairman 162.1 minister 161.7 delegation 158.9

53. @-@ 137.4 group 132.5 champion 128.0 **country 125.5** player 122.8 **team 119.8** member 117.1 victory 116.5 state 114.3 third 112.7

54. member 140.4 body 137.8 president 136.8 commission 113.3 committee 111.7 official 111.0 club 106.5 headquarter 98.1 federation 90.4 panel 82.6

55. victim 87.5 character 82.6 murderer 75.4 liar 63.4 traitor 63.4 bigot 63.4 criminal 63.4 fool 63.4 separatist 63.4 terrorist 63.4

56. change 180.6 style 179.0 thing 159.6 life 156.3 **quality 155.3** story 151.5 character 147.2 one 146.5 work 143.0 much 142.7

57. audience 75.4 success 69.9 criticism 63.4 **praise 63.4** attention 63.4 acclaim 63.4 **recognition 63.4** controversy 63.4 sale 63.4 money 63.4

58. breakthrough 131.3 effort 117.8 president 115.0 record 114.3 process 112.4 **achievement 108.5** result 108.2 victory 108.1 reform 104.8 measure 102.9

59. service 187.4 subject 187.4 development 181.9 study 179.7 art 172.7 **science 169.8** field 166.4 work 165.3 history 165.2 book 162.6

60. friend 151.8 lawyer 129.3 trial 102.3 president 94.4 brother 91.2 father 84.9 prosecution 80.6 associate 75.7 executive 75.4 attorney 67.5

61. product 160.0 source 139.5 component 137.5 substance 128.9 **chemical 127.4** plant 125.5 waste 113.8 food 111.5 ingredient 106.5 water 102.6

62. country 139.6 nation 127.5 summit 110.8 state 100.6 integration 98.4 **organization 95.5** meeting 87.0 cooperation 86.3 **grouping 86.1** foreign minister 85.8

63. house 206.7 area 206.4 family 180.6 street 153.8 car 135.8 market 128.1 block 126.8 field 106.4 lot 102.8 road 101.9

64. production 161.4 music 148.4 film 131.1 **musical 127.4 performance 125.6** adaptation 118.7 song 115.3 character 113.9 genre 113.3 script 112.2

65. star 119.0 **musician 106.5** singer 102.7 actress 101.4 actor 95.1 host 90.9 athlete 86.9 **music 85.6** troupe 84.0 dancer 79.9

66. title 161.5 event 154.5 champion 149.3 **tournament 137.5** championship 130.3 winner 126.8 major 126.4 final 114.8 open 97.8 first 92.8

67. team 176.1 game 159.3 club 122.5 record 101.6 hitter 97.7 series 91.4 pitcher 84.1 **baseball 82.9** title 82.1 card 80.3

68. politics 159.4 **religion 152.5** church 141.7 **belief 125.3** respect 122.0 work 120.8 faith 119.1 one 117.5 change 116.1 value 114.1

69. city 163.3 **town 144.0** capital 115.0 area 113.3 headquarter 97.7 province 96.0 stronghold 89.6 district 75.7 center 66.8 airport 58.3

70. number 121.6 name 113.0 key 97.8 information 87.1 site 85.5 list 84.0 code 75.4 date 69.9 card 69.8 map 67.5

71. paint 120.3 stain 109.4 **coating 106.4** chemical 91.5 color 89.3 material 86.4 coat 84.0 product 81.2 oil 77.7 surface 76.0

72. company 147.9 bank 120.5 investor 96.5 **shareholder 94.4** employee 91.3 executive 90.9 creditor 90.4 customer 90.4 one 85.1 money 84.2

73. **evidence 137.6** report 129.1 study 127.3 one 124.8 kind 123.3 experience 106.0 that 98.2 inquiry 98.1 discovery 97.7 reason 96.5

74. drug 124.9 substance 104.7 steroid 102.4 **hormone 94.4** effect 84.0 test 79.6 human growth hormone 63.4 testosterone 63.4 antibiotic 63.4 androstenedione 61.5

75. gainer 55.1 @@ to @@ 37.9 dow 35.4 jeff beck 32.6 laggard 32.6 issue 32.6 quarterback 32.6 versa 32.6 result 32.6 name 32.6

76. event 162.8 **sport 133.6 activity 101.9** swimming 90.4 trip 85.8 tour 85.8 competition 85.8 walking 78.9 basketball 78.9 men's 77.7

77. member 182.4 representative 177.3 meeting 172.6 session 166.9 hearing 153.7 official 147.1 **group 144.9** chairman 142. **committee 134.2** organization 129.2

78. **organization 63.7** china 63.4 world bank 63.4 government 63.4 head 57.3 expert 53.1 director-general 49.9 secretary-general 49.9 conference 49.9 general 49.9

79. **company 142.7** group 115.7 aerospatiale 63.4 british aerospace 63.4 joint venture 51.3 chairman 49.9 chief executive 49.9 share 49.9 unit 49.9 shareholder 43.0

80. clinton 110.2 investigation 108.6 matter 102.4 story 99.5 evidence 96.0 charge 96.0 **crime 94.4** failure 90.9 revelation 86.8 violation 85.7

81. people 78.9 middle class 63.4 home 56.8 folk 55.1 child 53.2 wealth 51.3 culture 47.0 more 45.0 blip 37.96 working class 32.6

82. mvp 63.4 cy young 63.4 selection 63.4 winner 60.4 runner-up 55.1 he 44.7 rookie 43.8 award 43.0 brett favre 42.2 trophy 40.5

83. game 120.5 period 110.6 time 106.2 deficit 83.9 first half 80.1 goal 76.1 @@@ million 74.3 one 72.2 lead 71.5 total 70.9

84. wine 140.8 **drink 127.2** product 117.5 sauce 106.4 **food 104.7** blend 99.5 beer 90.4 dish 89.9 ingredient 87.3 mixture 86.2

85. clergy 63.4 priest 63.4 leader 63.4 member 63.4 church 59.1 many 53.2 prayer 48.5 they 44.7 child 42.2 cathedral 36.2

86. **bank 105.9** branch 82.6 commercial bank 78.9 bank of china 63.4 industrial and commercial bank of china 63.4 department 60.1 shanghai branch 53.2 president 49.9 official 47.1 loan 46.2

87. production 128.7 music 116.3 film 114.1 **performance 90.5** giselle 85.8 **ballet 84.3** classic 77.7 song 74.9 movie 68.6 work 63.7

88. **coach 163.2** assistant 141.7 team 119.8 player 115.6 offense 108.2 friend 101.5 general manager 94.4 coordinator 85.0 patriot 77.7 he 74.9

89. child 214.9 one 209.4 son 207.9 daughter 207.0 family 206.6 wife 202.2 sister 199.3 father 198.7 parent 198.5 character 198.3

90. **weapon 126.0** system 114.5 warhead 106.4 **missile 104.1** launcher 85.8 hundred 83.1 vehicle 83.0 component 81.2 bomb 78.9 tank 78.9

91. fletcher energy 63.4 fletcher forests 63.4 air new zealand 63.4 fletcher paper 63.4 brierley investments 63.4 carter holt harvey 63.4 contact energy 63.4 auckland airport 63.4 fletcher building 63.4 fletcher forest 63.4

92. construction 106.4 building 105.1 area 97.3 passenger 77.8 **airport 76.5** center 74.1 bridge 70.6 feet 70.1 road 66.6 terminal 64.9

93. passenger 166.7 **ship 147.7** line 145.4 crew 143.7 **vessel 140.4** equipment 121.5 tank 117.8 weapon 112.2 **boat 106.0** platform 103.9

94. quarterback 132.3 pick 120.3 rookie 76.1 tim couch 63.4 akili smith 63.4 daunte culpepper 63.4 cade mcnown 63.4 ricky williams 63.4 mcnown 63.4 running back 63.4

95. game 165.2 race 141.2 second 139.5 third 133.1 **inning 126.1** team 124.9 double 122.18239095613913 sixth 120.3 ninth 110.3 fourth 106.9

96. sense 127.0 strength 111.5 vision 91.1 spirit 86.1 feeling 84.0 attitude 84.0 power 84.0 understanding 84.0 quality 78.9 humor 78.9

97. **sound 87.7** bell 78.9 light 78.9 blast 75.7 **noise 64.7** drum 63.4 horn 63.4 cheer 63.4 firework 63.4 alarm 63.4

98. **storm 120.3 hurricane 78.0** one 53.4 victim 49.9 brother 49.1 effect 46.2 wind 46.2 fear 43.0 coverage 43.0 forecast 37.9

99. **building 201.5** floor 178.3 room 173.4 door 168.0 complex 157.1 tower 155.8 project 154.3 space 153.2 center 144.6 window 144.1

100. vladimir stepania 63.4 vernon maxwell 63.4 horace grant 63.4 ruben patterson 63.4 brent barry 63.4 shammond williams 63.4 travis knight 32.6 each 32.6rashard lewis 30.7jelani mccoy 30.7