7-2011

Unsupervised Semantic Classification Methods

John Gilmer

Unsupervised Semantic Classification Methods

by

John Gilmer

Undergraduate honors thesis under the direction of

Dr. Jianhua Chen

Department of Computer Science

Submitted to the LSU Honors College in partial fulfillment of
the Upper Division Honors Program

Louisiana State University
Agricultural and Mechanical College
Baton Rouge, Louisiana

July 2011

Abstract

A current problem in text processing is the inability to make accurate unsupervised semantic classification systems. In this research we study the unsupervised semantic classification problem using several approaches. We find that morphological and semantic hints can be translated into effective rules within semantic classification. Our results showed a 66% recall rate and a 70% precision rate. We also observed that using raw contextual words as a metric for observing similarity between concepts is minimally effective. Finally we propose further research topics that may be able to improve recall and precision rates of unsupervised semantic classification systems.

Acknowledgements

I would like to thank Dr. Jianhua Chen for her patience, wisdom, and optimism. Dr. Chen's assistance was invaluable to the construction of this undergraduate thesis. I would also like to thank Dr. Coretta Douglas for her encouragement; she convinced me to pursue upper division honors.

Contents

Chapter 1

Introduction

Computational linguistics is a popular field of study within computer science. It encompasses all interacting processes between computers and natural languages. Among these processes are speech processing, text processing, translation, and knowledge extraction. Here we deal with text processing and knowledge extraction. This area of study has the potential to yield important results. If computers had the ability to understand or semi-understand natural language, then their influence in the realm of communication would be augmented.

1.1 Text Processing

Text processing refers to all operations a computer can perform on raw text. For instance, a computer could process a raw text and find grammatical errors or spelling errors in the natural language. This is currently done by programs like Microsoft Word. Another practical example is searching a raw text. There are database systems that are built to intelligently search raw text to process queries. These are just a few popular examples of successful text processing programs. Miller and Myers say that text processing is very important because there is a wealth of information stored within text [3]. There are many more possibilities that have not yet been realized. If a computer could read a raw text and build a knowledge structure based on the document, then there would be many practical applications. For instance, a computer could process a complicated legal document and answer simple queries regarding the document. Complete natural language understanding

by a computer will not be achieved for a very long time, if ever. Nevertheless, computer scientists and linguists are working towards achieving better understanding by many different means. One important area of research is semantic classification. With semantic understanding, a computer can begin to construct a knowledge structure.

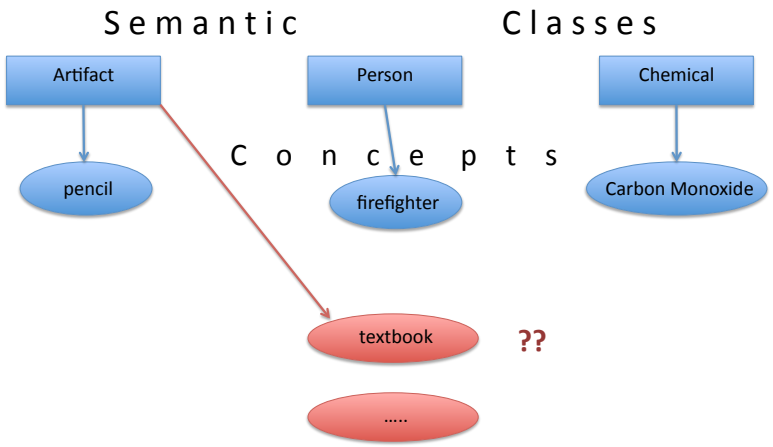
1.2 Semantic Classification

Semantic classification refers to the classification of concepts into different semantic categories. For instance, the word "pencil" would be classified in the semantic category of *artifact* because it is an object constructed by humans. A concept can be classified into any number of different semantic classes, and in fact it can sometimes belong to more than one class. In order for a computer to build knowledge from text, semantic understanding is essential. If a computer can acquire the semantic knowledge that a "firefighter" is a person, then it also can inherit some knowledge about people in general and apply it to the "firefighter". The computer would then know that "firefighter" is a person and so has some properties like family, job, possession, and body parts. Figure 1.1 displays the structure of semantic classification.

Semantic knowledge is important in ontology construction. Ontologies are an emulation of a domain. An ontology is better described by its composition than a definition. Ontologies are composed of concepts, attributes, taxonomy, and non-taxonomic relations. The most important area in our research is taxonomy. A taxonomy is a hierarchical structures that is used for organizational purposes. An example of a taxonomy is the scientific classification of different species. The four kingdoms of life each have subclasses within them like phylum, class, order, family, genus, and species. Taxonomies can be used to define or organize any set of information. In the case of natural language, computer scientists want to make a semantic taxonomy. For instance the word "baseball" may be within the category *sport* which is within the category *game* and so on and so on. In our research we attempt to build a small taxonomy that is only two levels deep by assigning concepts to semantic classes. Figure 1.2 depicts the general process of semantic classification and how it builds a taxonomy.

This research paper explores the question of how to make a computer accurately classify concepts. There is a lot of current research in this area; Fan and Friedman [2] explore semantic classification of biomedical concepts

Figure 1.1: Semantic Classification Example

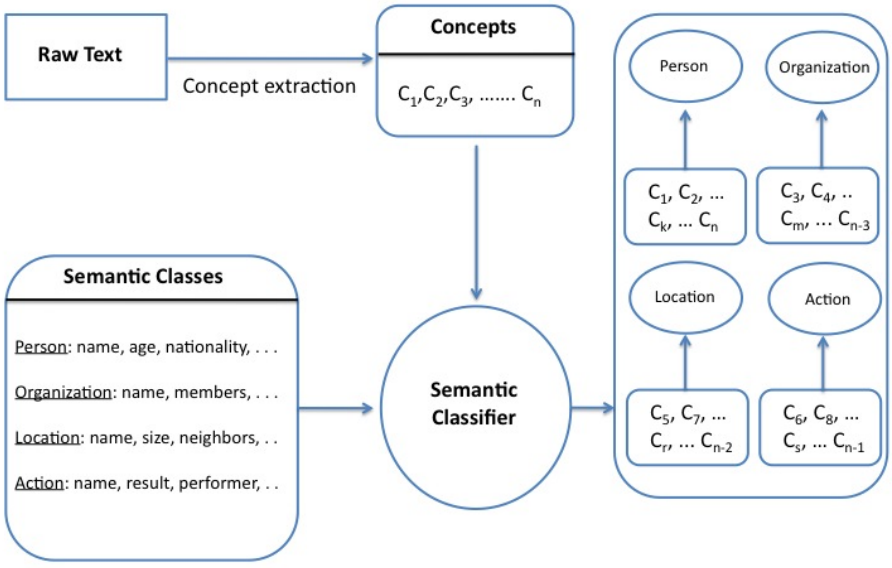


for the purposes of building biomedical semantic knowledge. It is relatively simple for a human to classify a concept. We know that "Iran" is a location because we have that knowledge in our mind. A computer would encounter the concept "Iran" and have no knowledge of what semantic class it may belong to. So how can semantic knowledge and understanding be built into a computer so that it can classify different concepts? There have been many studies in this field to find the best ways to accomplish this important task. They fall into two categories: supervised and unsupervised semantic classification.

1.3 Supervised Semantic Classification

Any supervised learning problem requires massive amounts of training data and is domain dependent. Supervised semantic classification is usually domain dependent as the training data is built from the domain itself. This is a problem because assembling training data is work intensive. The accuracy

Figure 1.2: SCL Framework



of supervised semantic classification tends to be better than unsupervised semantic classification. The accuracy of supervised semantic classification is very desirable, but the amount of tedious work required is undesirable. A better method would be accurate unsupervised semantic classification.

1.4 Unsupervised Semantic Classification

Unsupervised semantic classification requires no training data and is domain independent. The computer processes the text looking for concepts. When the computer encounters a concept, it attempts to classify it into one of the predefined semantic categories based on its similarity to concepts already placed into the semantic categories. This method requires some initial seeds to populate the semantic classes so that the first concept classified will have other concepts to compare with. The metrics of comparison vary with different approaches.

1.5 Problem and Approach

Judging from the current research climate in the area, it seemed important to investigate unsupervised semantic classification and how improvements in this area could be made. To investigate this problem we tried using many different types of properties to determine similarity between concepts. We evaluated the effectiveness of contextual words, morphological heuristics, and part-of-speech properties in classifying concepts with unsupervised learning.

1.6 Summary

Building a knowledge structure from semantic information is tantamount to approaching natural language understanding for a computer. Semantic classification plays an important role in this process, and if unsupervised semantic classification can be done accurately, then building semantic knowledge structures will become both easier and more reliable. It is for these reasons that the investigations into unsupervised semantic classification are an important part of computer science research.

Chapter 2

Background

Before describing our methods of investigation, it is necessary to introduce previous research in semantic classification and discuss how our experiments were inspired from the results of previous research. Punuru's PhD thesis at LSU investigated both supervised and unsupervised semantic classification [4] [5]. Both of these experiments influenced our decisions in implementing this unsupervised semantic classifier. Additionally, a research paper on classification of proper nouns proved to be influential [8]. Both of these reports were important in designing our experiments because our concept set differed from Punuru's in that we did not ignore proper nouns but instead included proper nouns and non-proper nouns.

2.1 Punuru's Supervised Semantic Classification

Punuru selected 4 different semantic categories in which to place concepts that were extracted from his text domain. The domain is from *New York Times* articles about electronic voting. The semantic categories that were used were *Person, Artifact, Location, and Action*. Punuru used a naive Bayes classifier with four different attributes for each concept to distinguish between semantic classes. The attributes were:

1. Last two characters of the concept.

2. Headword of the concept.

3. Pronoun following the concept.

4. Preposition preceding the concept.

All of the training data is collected and each concept in the training data is assigned data for each of the four properties. During processing, each of these attributes is collected for every single concept that is extracted. The concepts are then applied to the Naive Bayes classifier to determine which semantic class they are closest to, according to their similarity to the members of those classes. To better illustrate how these four properties work, a table is supplied.

Table 1: Training Instances

Concept	Attribute Values	Class
intimidation	on, intimidation, NPRN, NPREP	Action
precinct	ct, precinct, their, to	Location
machine	ne, machine, it, NPREP	Artifact
keyboard	rd, keyboard, its, to	Artifact
governor	or, governor, it , NPREP	Person
prison	on, prison, their, NPREP	Location
manipulate	te, manipulate, their, NPREP	Action
country	ty, country, NPRN, in	Location
resident	nt, resident, we, NPREP	Person

The results of this supervised semantic classification were quite impressive. The Naive Bayes classifier, with these four attributes, produced an average accuracy of 93.6%. In further analysis each attribute was evaluated for its efficacy by seeing how the accuracy results changed when certain attributes were not used. By this metric it seemed that the last two letters of the word attribute and the attribute for preposition preceding the concept were not as effective as the other two attributes [4]. After reviewing this part

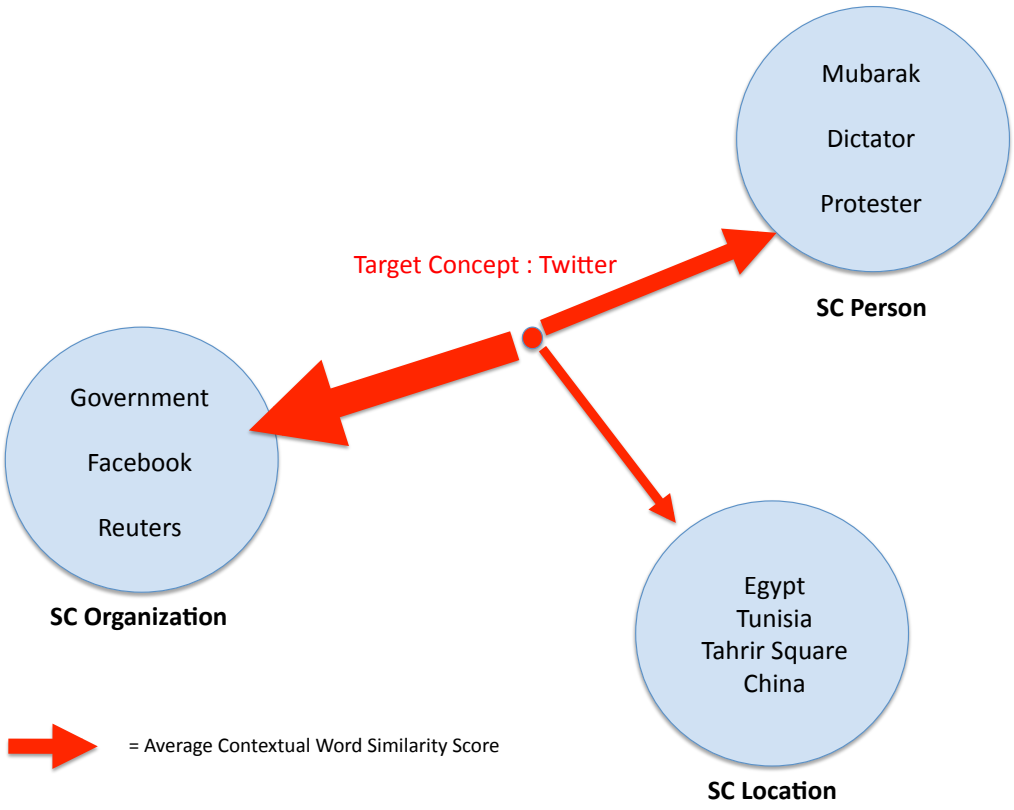
of Punuru's research, we thought that the headword attribute would prove to be effective in classifying concepts.

2.2 Punuru's Unsupervised Semantic Classification

For unsupervised semantic classification, Punuru used completely different attributes. Rather than taking knowledge from the concept itself, he classified concepts by looking at their contextual words [4]. Unsupervised semantic classification does not involve training data. Instead each semantic category is given three or four *seed* concepts. These seeds define the semantic category. Each seed concept and each target concept are represented by the contextual words that surround them. Similarity between two concepts is based on similarity between their contextual words. When the computer tries to classify a concept, it looks for the semantic category that the concept is most similar to, based on contextual words, and then adds the target concept to that group. Figure 2.1 displays the process of classification by a similarity metric.

The process of assigning each new concept that the computer recognizes is known as *snowballing* because the semantic classes get larger with each assignment. The details of this algorithm are displayed in the Initial Approach chapter. The results of unsupervised semantic classification paled in comparison to the accuracy of the supervised semantic classification. The best results were reached with a similarity threshold of .15 [4]. This means that if the similarity score between two concepts was less than .15, then it was ignored. Using this threshold, the program was able to classify only concepts it was confident about. The results of this work were a recall measurement of 24% (24% of the target concepts were classified) and a precision measurement of 73%. These numbers are far off from the 93.6% accuracy of the supervised semantic classification method. Punuru speculates that there are more suitable features to be used for unsupervised semantic classification [4]. In our experiment we would try using different features and also using different snowballing algorithms that may be more accurate.

Figure 2.1: Semantic Classification By Similarity



2.3 Proper Noun Classification

Smarr and Manning discovered some interesting results about classifying proper nouns [8]. They concentrated solely on classifying proper noun phrases into 5 different categories: *drug names, company names, movie titles, place names, people's names.* Their learning system works in a domain independent way without building in any heuristics. Their inspiration was that many proper nouns were easy to place by humans simply by looking at them. For instance the word "Novo-Doxylin" just looks like it would be a drug [8]. They proposed that the word itself provides sufficient information to classify it in a probabilistic manner. With a probabilistic model based on word length

and character sequences they were able to achieve high levels of accuracy in classifying proper nouns. The idea that the composition of the word itself can give hints about its semantic category is interesting, and it was used in our research.

2.4 Summary

The strategies from these research papers inspired us to try using many different properties of concepts in hope to accurately classify them with unsupervised learning. Because our concept set consisted of both proper and non-proper nouns, both strategies from these research papers proved useful. Our investigations included studying the effectiveness of alternate clustering methods, using part-of-speech data, using contextual information, and using particular character strings in classifying a set of target concepts. In the next chapter we discuss the domain, concept extraction, target concepts, algorithms, and attributes in more depth.

Chapter 3

Initial Approach

Our methods for this experiment in some ways mirrored Punuru's method for the sake of comparison. His unsupervised results were quantitatively represented as 24% recall and 73% accuracy. With these results in mind, if we approach the problem in a similar fashion then we can make a more accurate conclusion about the effectiveness of our experiments.

3.1 Domain

In order to perform meaningful experiments we first require a meaningful domain. The domain must include only texts relevant to a single subject. The more specific the domain is, the better classification will be because the text will involve a high frequency of similar concepts. The domain for this project was a collection of 19 articles produced by the *New York Times*. These articles were selected for their association with the concept of the "Facebook Revolutions." The "Facebook Revolutions" is a term that refers to the series of revolutions sweeping across the Arab world in the spring of 2011. This domain would eventually generate a concept set with the some of the most frequent concepts being "revolution," "Mubarak," and "Egypt." For a list of concepts see Appendix B.

3.2 Concept Extraction

Acquiring an appropriate set of concepts from text is a distinct area in Text Mining. This area is called Concept Extraction. There are different processes

and strategies for acquiring the best set of concepts. There is one necessary process that must occur before a set of concepts can be acquired, and that is the process of part-of-speech tagging. Part-of-speech tagging takes a document of natural language and assigns to each word a part-of-speech indicator. For instance an adjective like "green" would be assigned the adjective tag "JJ." There is a large list of the different part of speech tags [6], and these are located in Appendix A.

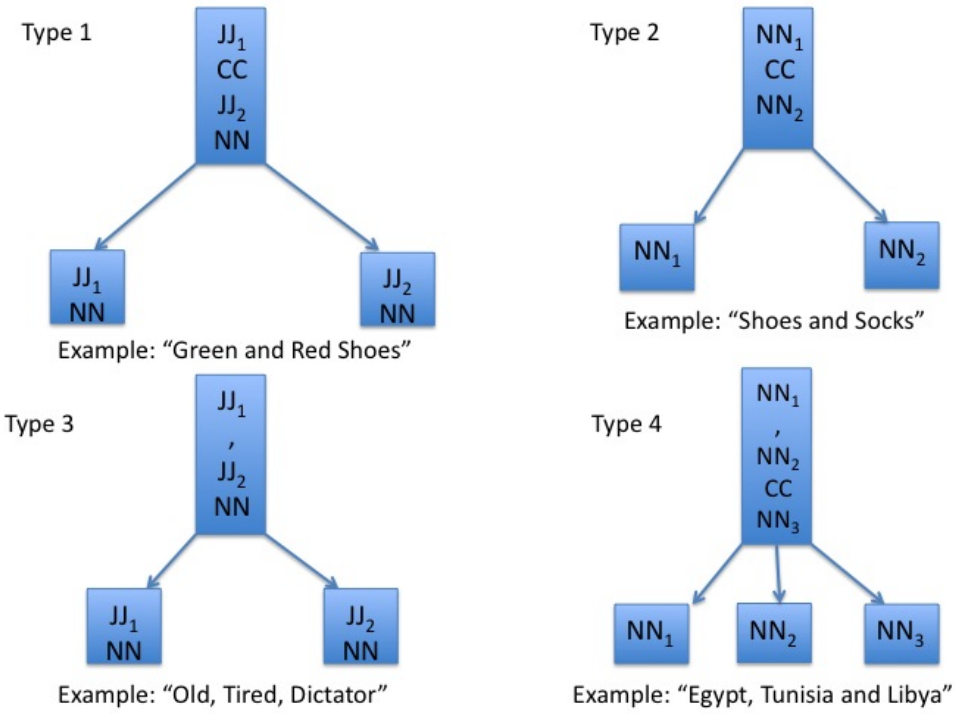
Part-of-speech tagging is the first step in concept extraction. There are many different POS (part-of-speech) taggers; the one we used was developed at the Institute for Computational Linguistics of the University of Stuttgart [7]. This POS tagger was developed by Helmut Schmid. The tagger has 96% accuracy, so errors are negligible. In addition to tagging words, Helmut Schmid's program also chunks phrases. Chunking is an important part of concept extraction. Chunking places marks in the text document where noun phrases or verb phrases begin and end. Using Helmut Schmid's tagger and chunker we were able to extract a large set of concepts. We placed every noun phrase into the concept set. This yielded some repetitive and erroneous results. Further concept-extraction processes were required to attain a quality concept set.

There are some words that occur too frequently in regular English to have any effect on classification. For these words it is best to ignore them. These words are referred to as "stop words," and ignoring stop words was the first strategy in improving the concept set. Any concept that began with a stop word was reset to the same concept without the first word, and any concept that was composed of solely stop words was ignored entirely. Our list of stop words consists of about 150 words with words like "then" or "he" or "at." Removing stop words is the first method of improving the concept list; the next method deals with morphological considerations.

There are many concepts that essentially have the same meaning but are morphologically different; to improve the concept list it is best to ignore all of the morphological differences. For instance, here are two separate concepts that should be considered a single concept: "Egyptian" and "Egyptians." These redundancies are eliminated by using the root form of all words that are plural (indicated by the POS tag). Another morphological disturbance is caused by the possessive. This problem is solved in the same way as the plural. Removing morphological redundancies improves the concept list again; the last improvement we made on the concept list was to separate conjunctive concepts.

Conjunctive concepts involve CCs (coordinating conjunctions) and/or commas. These concepts are a combination of multiple separate concepts. For example "Egypt, Tunisia, and Libya" is a separate concept from "Egypt." The last improvement to the concept set was to separate some conjunctive concepts into their separate parts. Figure 3.1 describes the different types of conjunctive phrases.

Figure 3.1: Types of Conjunctive Phrases



Type 1 and Type 3 should not be separated because both adjectives make up the entire concept. Type 2 and Type 4 should be separated with each comma and with each CC. While processing the concepts our program looked to make sure a separate noun phrase existed before each CC and before each comma before separating it into its own concept. So, after removing stop words, adjusting morphological differences, and separating conjunctive phrases, we reach our final concept set. The process of concept extraction is best expressed by the following table:

Table 2: Concept Extraction

Step Name	Output
Raw Text	Today the campaign has grown, complete with a Web site commemorating victims, an active Facebook page and over 1,000 followers on Twitter. Despite the cost of personal computers and connectivity, the number of Facebook users in the Middle East and North Africa is increasing rapidly.
POS Tagging	Today/NN the/DT campaign/NN has/VBZ grown/VBN, complete/JJ with/IN a/DT Web/NP site/NN commemorating/VBG victims/NNS, an/DT active/JJ Facebook/NN pageNN and/CC over/IN 1,000/CD followers/NNS on/IN Twitter/NP. Despite/IN the/DT cost/NN of/IN personal/JJ computers/NNS and/CC connectivity/NN, the/DT number/NN of/IN Facebook/NN users/NNS in/IN the/DT Middle/NP East/NP and/CC North/NP Africa/NP is/VBZ increasing/VBG rapidly/RB.
NP Chunking	[Today/NN] [the/DT campaign/NN] has/VBZ grown/VBN, complete/JJ with/IN [a/DT Web/NP site/NN] commemorating/VBG [victims/NNS], [an/DT active/JJ Facebook/NN pageNN] and/CC over/IN [1,000/CD followers/NNS] on/IN [Twitter/NP]. Despite/IN [the/DT cost/NN] of/IN [personal/JJ computers/NNS and/CC connectivity/NN], [the/DT number/NN] of/IN [Facebook/NN users/NNS] in/IN [the/DT Middle/NP East/NP and/CC North/NP Africa/NP] is/VBZ increasing/VBG rapidly/RB.

Step Name	Output
Stopword Processing	Today/NN, campaign/NN, Web/NP site/NN, victims/NNS, active/JJ Facebook/NN pageNN, 1,000/CD followers/NNS, Twitter/NP, cost/NN, personal/JJ computers/NNS and/CC connectivity/NN, number/NN, Facebook/NN users/NNS, Middle/NP East/NP and/CC North/NP Africa/NP
Morphological Analysis	Today, campaign, Web site, victim, active Facebook page, 1,000 follower, Twitter, cost, personal computer and connectivity, number, Facebook user, Middle East and North Africa
Conjuction Separation	Today, campaign, Web site, victim, active Facebook page, 1,000 follower, Twitter, cost, personal computer, connectivity, number, Facebook user, Middle East, North Africa

3.3 Concepts

After tagging and chunking all 19 documents, we had a preliminary concept set. Then after some concept-extraction procedures, we had a better concept set without erroneous or redundant entries. Our final concept set had 3610 different concepts. By inspecting the most frequently occurring concepts it became clear that not many of the concepts belonged to the semantic class *artifact* so we decided to remove that semantic class and replace it with the semantic class *organization*. To test our unsupervised semantic classification we selected 75 concepts belonging to each of the 4 semantic classes. These 300 concepts, ordered by their frequency of occurrence within the domain, made up the set of target concepts. They are located in Appendix B.

3.4 Similarity Metric and Algorithm

Here we discuss the basic program we developed. In the Extended Approach Chapter we will discuss the different variations we applied in order to study the usage of other attributes besides contextual words. The basic unsupervised semantic classification method is based solely on contextual words. Each of the 3610 concepts has a vector of contextual words with frequency data also. When a concept is collected, contextual words are assigned to its

vector, two from the left and two from the right. These vectors keep track of which concepts have which contextual words and they keep track of the frequency of the contextual words. In this example of a vector, we have the first instance of a concept followed by its vector of contextual words, each contextual word is in the root form of the original word and has a frequency of 1 because it is the first time this concept is encountered. If the concept "Arab World" is encountered again, then its frequency will increment and new contextual words will be added or incremented.

Concept: Arab World , Frequency = 1
Contextual Word Vector: <rotten, tackle, help, homegrown>

All of this information is used in calculating mutual information and similarity. The computer assigns a target concept to a semantic class based on its similarity to the other concepts within that class, but similarity must be quantifiable; that is what the vectors are for. Along with the contextual word vector, each concept has a mutual information vector that has 3921 entries (the size of the total number of distinct contextual words). This vector is calculated as follows:

$$MI(C_i, f_j) = \frac{C(C_i, f_j)}{C(C_i) * C(f_j)}$$

MI = Mutual Information function. $C = C_1, C_2, ..., C_n$ is the set of concepts. $F = f_1, f_2, ..., f_m$ is the set of all the contextual words. $C(C_i, f_j)$ is the count of occurrence of contextual word f_j with concept C_i. $C(C_i)$ is the total count of frequency of C_i in the domain, and $C(f_j)$ is the count of the frequency of f_j as a contextual word to any concept. The Mutual Information vector is calculated for every concept immediately after the program has collected all concepts and contextual words. Then, after the seeds have been set, the computer begins assigning target concepts to the proper semantic class. The semantic classes are represented by the set $SC = SC_1, SC_2, SC_3, SC_4$. The algorithm for this unsupervised semantic classification follows:

1. initialize each SC_j with seeds
2. for each concept C_i in C
   for each SC_j in SC
   $$avg_j = \frac{\sum_{C_k \in SC_j} Sim(C_k, C_i)}{|SC_j|}$$
   $SCI = \arg\, max_{1 \leq j \leq 4} avg_j$
   $C_i \in SC_{SCI}$

The function $Sim(C_k, C_i)$ is the cosine similarity function. This algorithm goes through the set of target concepts and assigns them to the semantic class that they best match using the mutual information vector as a metric. This algorithm is the basic concept of unsupervised semantic classification using contextual words. This type of learning that builds on itself is sometimes known as *snowballing*; we will use this term in the following chapters.

3.5 Summary

In this chapter we have described our data and detailed our methods of retrieving the data. In addition we have shown the algorithms used to create the snowball effect. After establishing the methods of our project we can now describe the results. In the Extended Approach chapter we will discuss the effectiveness of the basic unsupervised semantic classification based on contextual words, and then we will describe how using other properties and using different snowball algorithms can affect the success of unsupervised semantic classification.

Chapter 4

Extended Approach and Results

In this chapter we discuss the results of the basic unsupervised semantic classification method using contextual words; then we present other variations of this method. The purpose of this study is to examine the strength of contextual words and other types of information in classifying concepts.

4.1 Frequency-Ordered Unsupervised SCL

In the basic method described in the last chapter, the set of target concepts is ordered by frequency of the concept. This means that the target concepts are processed and assigned in the order of their frequency. The table below shows the results of this frequency-ordered unsupervised SCL (semantic classification):

Table 3: Frequency-Ordered Unsupervised SCL

Correct	Incorrect	Not Placed	Recall	Precision
30%	70%	0%	100%	30%

This table shows that only 30% of the target concepts were assigned correctly, and the rest were assigned incorrectly. Immediately it seems that contextual words are not so helpful in classifying concepts. If we were to assign concepts randomly to each semantic class then we would have a 25% precision rate. So, using contextual words as data for unsupervised learning only increased the precision rate by 20%. It is possible that the contextual

words are helpful, but we are not using the data properly. Instead of ordering the target concept set by frequency we could order it in a different manner so that the program learns better.

4.2 Order by Confidence Ratio

The results of the frequency-ordered unsupervised SCL did not show good precision. Our hypothesis was that this was because the ordering of the target concepts was not optimal. We thought that we could order the target concepts in such a way that the best-fitting target concept would be assigned first. This way the snowball effect would not get out of control. In order for a target concept to be assigned to a semantic class the program goes through all of the target concepts and gets each concepts' average similarity scores for each of the semantic classes. The confidence ratio is calculated for each concept, and the concept with the highest confidence ratio is assigned to its appropriate semantic class; then the program goes through the rest of the target concepts in the same manner. This classification algorithm runs much slower than the frequency-ordered one; it runs at $O(n^2)$ where n is the number of target concepts. The Confidence Ratio is defined below:

$$CR = \frac{avg_1}{avg_2}$$

$$avg_1 = \text{largest average score}$$
$$avg_2 = \text{second largest average score}$$

Our results were similar to the frequency-ordered unsupervised SCL as is seen in the table below:

Table 4: Confidence Ratio Unsupervised SCL

Correct	Incorrect	Not Placed	Recall	Precision
25%	75%	0%	100%	25%

These results again indicate that the contextual words do not provide sufficient information to classify concepts with an unsupervised learning method. Although the confidence ratio did not show an improvement, thresholding may be able to increase the precision rate.

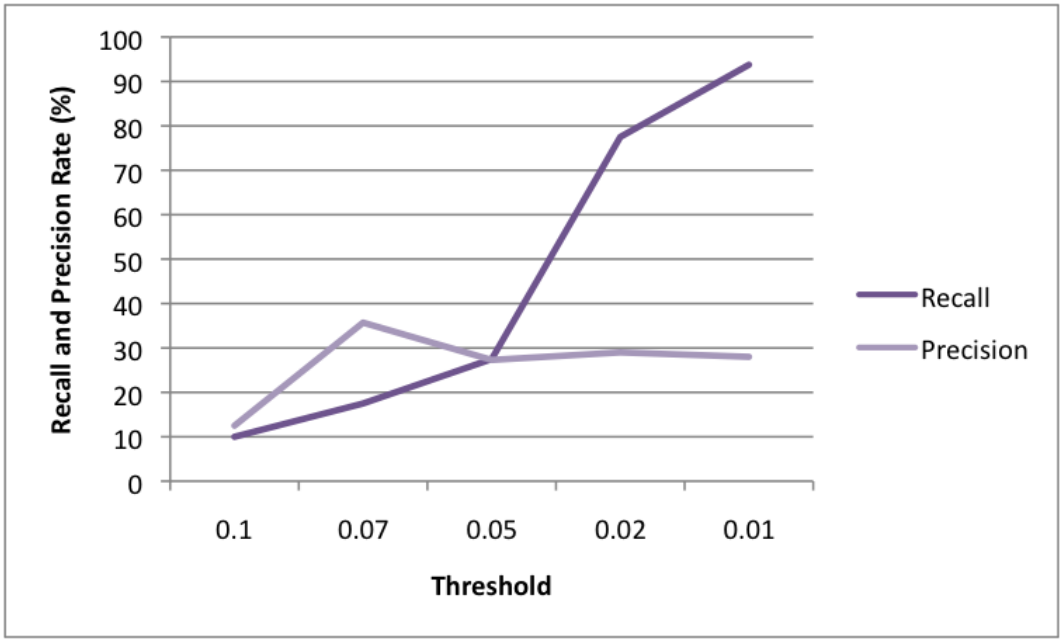
4.3 Thresholding

Thresholding is the process by which target concepts can be ignored rather than being assigned to a semantic class if their average similarity score is too low to be considered relevant. We did some experiments with thresholding on the frequency-ordered unsupervised SCL to see if we could improve the precision rate. A table, along with figure 4.1, is shown below:

Table 5: Thesholding for Frequency-Ordered Unsupervised SCL

Threshold	Correct	Incorrect	Not Placed	Recall	Precision
.1	1.25%	8.75%	90.0%	10.0%	12.5%
.07	6.25%	11.2%	82.5%	17.5%	35.7%
.05	7.5%	20.0%	72.5%	27.5%	27.3%
.02	22.5%	55.0%	22.5%	77.5%	29.0%
.01	26.25%	67.5%	6.25%	93.75%	28.0%

Figure 4.1: Threshold Evaluation Graph



In this table it is seen that with a threshold of .07 a recall rate of 17.5% and a precision rate of 35.7% are achieved. Although the recall rate is very

low, the precision did see an increase of approximately 17% from the basic frequency-ordered unsupervised SCL. In summary, this table and the results of the confidence ratio SCL show that using raw contextual words as data is only slightly helpful in classifying concepts in an independent domain. Next we investigate how heuristics, including semantic and morphological clues, can help to classify concepts.

4.4 Heuristic Approach

Smarr and Manning [8] had shown earlier that the morphological structure of a word, especially a proper noun, is very powerful knowledge in assigning concepts to semantic classes. Additionally, Cuchiarelli [1] showed in his study of unknown proper nouns that some semantic heuristics are very helpful, for instance that some precursors ("Mr." or "Mrs.") indicate which semantic class the concept belongs to. Without using a probabilistic model like Smarr and Manning, we formed a domain-independent set of morphological and semantic rules to create a more accurate unsupervised SCL. These rules do not enforce assignment of concepts; they only influence the scores of each concept. The weights used for increasing the various class scores (50,100, 1000) are determined empirically. The list of rules can be seen in Table 6.

These heuristic rules help to classify some of the concepts in the target concept set. We used these heuristics to make more informed classifications, but we also used them as a threshold. If a concept's scores were in no way affected by the heuristic rules, then we did not assign it to any semantic class. If a concept is not a proper noun and does not have any of the appropriate suffixes of non-proper nouns, then the concept will be ignored and not placed into a semantic class. Table 7 shows the results achieved by using the threshold and by not using the threshold.

Table 6: Heuristic Rules

Legend: TC = Target Concept, PN = Proper Noun, OS = Organization Score, AS = Action Score, PS = People Score, LS = Location Score	
Test	Action
If TC is a PN	
If TC has a '.' in the 2nd or 3rd place	PS = PS * 1000 AS = 0
If TC has suffix -a or -o	LS = LS * 1000 AS, OS = 0
If TC is not a PN	
If TC has suffix -ment or -sion	AS = AS * 50 PS, LS = 0
If TC has suffix -tion	AS = AS * 50 PS = 0
If TC has suffix -ence or -al	OS = 0
If TC has suffix -ant, -ent, -yst, -ist, -er or -or	PS = PS * 1000
If TC has suffix -y	OS = OS * 10 LS = 0

Table 7: Heuristic Rules SCL Results

Threshold	Correct	Incorrect	Not Placed	Recall	Precision
Yes	46.7%	19.3%	34.0%	66.0%	70.7%
No	56.1%	43.9%	0.0%	100.0%	56.1%

These results are quite good and compared well with Punuru's results. Punuru achieved 24% recall and 73% precision. Our most accurate results showed 66% recall and 70.7% precision. The precisions are about equivalent and our recall was nearly 3 times greater. The real surprising information is that even when the semantic classes have been filled with approximately 138 correct concepts and only 60 incorrect concepts, the remaining target concepts are not assigned very accurately. This is seen in the chart where the precision decreases when the concepts without heuristic rules are assigned based solely on contextual words (the "No" Threshold).

There is a statistical metric used to compare results that combines both the recall and precision rates. This metric is called the F-measure; it is a

harmonic mean of the recall and precision rates. We will use the F-measure to compare our results with a single value. The F-measure equation is given below along with a table comparing our results via the F-measure statistic:

$$F = 2 * \frac{precision * recall}{precision + recall}$$

Table 8: F-measure

System	F-measure
Punuru's Unsupervised SCL	.361
Heuristic Approach (Yes Threshold)	.683
Heuristic Approach (No Threshold)	.719

The F-measure table shows that the results we achieved with the heuristic approach surpassed the results that Punuru had achieved in his unsupervised semantic classification system that only used contextual words as data.

4.5 Summary

In order to compare our results to Punuru's we should again restate that Punuru used a concept set without proper nouns while our concept set included both proper and non-proper nouns. Perhaps these differences between concepts are responsible for the variation of results of unsupervised semantic classification. These results lean ostensibly towards the strength of morphological hints and away from contextual words. We will inspect these ideas further in the Conclusion Chapter.

Chapter 5

Conclusion

Our results have shown that contextual words are minimally helpful in unsupervised semantic classification while morphological and semantic hints are extremely useful. Our snowball method combined with heuristic rules applied to a full concept set (proper nouns and non-proper nouns) achieved a result that improved on similar previous research. With the application of heuristic rules we managed to improve Punuru's unsupervised semantic classification system from 24% recall and 73% precision to 66% recall and 70.7% precision.

It is natural to wonder why contextual words do not work so well in their current format. Similarity between concepts can only be generated if the exact same contextual words occur in the context of the two concepts. This does not account for synonyms or similar words. A possible solution would be to check the WordNet inherited hypernym hierarchy. WordNet is a lexical database of English produced at Princeton University. For instance the words "help" and "aid" are very similar, and their first inherited hypernyms are both "activity" while "dinosaur" and "rock" have different inherited hypernyms. The distance between matching inherited hypernyms could be used as a metric to evaluate contextual word similarity. This solution may produce better results rather than simply relying on raw contextual word matching. This solution would not improve proper noun classification and for that we believe that studying morphological and semantic hints may lead to very successful unsupervised semantic classification systems.

We have identified two important areas of study that offer promising results. With future research, unsupervised semantic classification systems should see vast improvement. With accurate semantic classification systems

that are domain-independent and do not require large sets of training data, text processing becomes much more powerful. Computers will be able to build knowledge for specific domains and will begin to understand natural language better. We hope that our research in this area will encourage other researchers to pursue improvements in unsupervised semantic classification.

Bibliography

[1] Cuchiarelli, A., Luzi, D., and Velardi P. (1999). Semantic Tagging of Unknown Proper Nouns. *Natural Language Engineering* (pp. 171-185).

[2] Fan J. and Friendman, C. (2007). Semantic Classification of Biomedical Concepts Using Distributional Similarity. In *Journal of the American Medical Informatics Association*. Volume 14, Issue 4.

[3] Miller, R. and Myers, B. (1999). Lightweight Structured Text Processing. In *Proceedings of the USENIX Annual Technical Conference*.

[4] Punuru J. (2007). Knowledge-Based Methods for Automatic Extraction of Domain-specific Ontologies. Ph.D. dissertation, Louisiana State University, 2007.

[5] Chen, J. and Punuru, J. (2007). Learning for Semantic Classification of Conceptual Terms. In *Proc. of of IEEE International Conference on Granular Computing 2007*.

[6] Santorini, B. (1991). Part-of-Speech Tagging Guidelines for the Penn Treebank Project.

[7] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*.

[8] Smarr, J. and Manning, C. (2002). Classifying Unknown Proper Noun Phrases Without Context. Technical Report. Stanford

Appendix A

Part of Speech Tags

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NP	Proper noun, singular
15.	NPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PP	Personal pronoun
19.	PP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle

24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Appendix B

Target Concepts

Figure B.1: Location Target Concepts

Egypt	Tripoli	Office
Country	West	Tahrir
Tunisia	Prison	America
Tahrir Square	Tobruk	Hospital
Street	Israel	Headquarters
Cairo	Café	Switzerland
World	Amman	Davos
China	United States	Ajdabiya
Arab World	Village	Islamic Country
Region	Syria	Djibouti
Libya	Belarus	Area
Middle East	Workshop	Afghanistan
Bahrain	Benghazi	Iraq
Jordan	Border	Cairo Slum
Town	Algeria	Cathedral
Home	House	France
Place	Saudi Arabia	Nile Delta
State	India	Islamic State
Yemen	East	Lobby
Capital	Beijing	Station
Square	Tunis	Barracks
City	Myanmar	Police Station
Iran	Europe	Damascus
Bulgaria	Russia	Club
Morocco	Venezuela	California

Figure B.2: People Target Concepts

People	Peer	Director
Mubarak	Young Protester	Foreigner
Mr. Sharp	Elite	Aide
Mr. Mubarak	Many Egyptian	Witness
Young People	Young Man	Resident
Member	General	King
Man	President Hosni Mubarak	Cousin
Colonel Qaddafi	Autocrat	Opponent
Everyone	Parent	Prisoner
Woman	Founder	Mr. El-Erian
Egyptian	Professor	Student
Mr. Maskati	Bouazizi	Police Officer
Leader	Hosni Mubarak	Somebody
Mohyeldin	Morozov	Democrat
Family	Reporter	Mr. Durbin
Demonstrator	Son	Wael Ghonim
Dictator	Dissident	Civilian
Crowd	User	Officer
Youth	Mr. Qaradawi	Security Official
Anyone	Secretary	Mr. Suleiman
Analyst	Governor	Col. Muammar El-Qaddafi
President	Young Woman	Soldier
Father	Hafez	Mohamed ElBaradei
Citizen	Friend	Sister
Activist	Someone	Mr. Nemtsov

Figure B.3: Action Target Concepts

Revolution	Attempt	Economic Activity
Protest	Return	Social Justice
Uprising	Concession	Warning
Violence	Peaceful Protest	First Protest
Change	Political Change	Aggression
Movement	Violation	Decision
Demonstration	Police Abuse	Formation
Use	Celebration	Arab Uprising
Work	Transition	Arrival
Talk	Attack	Confrontation
Study	Foreign Investment	Economic Growth
Struggle	Isolation	Reconciliation
Liberation	Coverage	Oppression
Repression	Insurrection	Pledge
Clash	Enrichment	Abuse
Call	Libyan Revolt	Brutality
Debate	Protest Movement	Parliamentary Election
Unrest	Investment	Tunisian Revolution
Humiliation	Kickback	Experiment
Response	Development	Argument
Discussion	Toppling	Behavior
Election	Sea Change	Binge Drinking
War	Achievement	Small Donation
Nonviolence	Arab Revolution	Observation
Update	Departure	Presidential Election

Figure B.4: Organization Target Concepts

Facebook	YouTube	Cabinet
Government	State Television	Egyptian Society
Group	Authoritarian International	Civil Society
Brotherhood	Business	System
Police	Egyptian Government	Reuters
Twitter	Council	Mubarak Government
Internet	Monarchy	Shiite Majority
Party	Network	Authoritarian Rule
Society	Campaign	Committee
Company	Facebook Page	Governing Party
Economy	Political Organization	Transitional Government
Army	Authoritarian Regime	CNN
Islamic Group	Web Site	Opposition Group
Military	Former Regime	Community Project
Security Force	Security	Armed Forces
Al Jazeera	Harvard	Political Party
Organization	Social Network	Repressive Government
Dictatorship	Associated Press	Web
University	Staff	Facebook Group
Regime	Department	Despotism
Supreme Council	Washington	Western Government
New Government	Coalition	Mubarak Regime
Google	Global Economy	Doha Film Institute
Old Regime	Religion	Online Community
Parliament	Small Group	Community