Using Semantic Sets for Word Sense Disambiguation and Keyword Extraction

Ganesh Ramakrishnan
Dept. of Comp Science and Engg.
Indian Institute of Technology,
Mumbai 400076,
Maharashtra, India.
Email: *hare@cse.iitb.ac.in*

Pushpak Bhattacharyya
Dept. of Comp Science and Engg.
Indian Institute of Technology,
Mumbai 400076,
Maharashtra, India.
Email: *pb@cse.iitb.ac.in*

Abstract

This paper presents an automatic method for resolving the lexical ambiguity of words in any free-flowing text, by using the taxonomy present in the WordNet and also the relative position of words in the given text, to construct semantic sets from the text. The semantic set has been defined as a collection of senses of words in given text that are related in the WordNet hypergraph structure. Two natural consquences of recognising semantic sets are *word sense disambiguation*, and *keyphrase extraction*. The method is supported by good empirical results.

1 Introduction

1.1 Word Sense Disambiguation

Any language uses words with multiple meanings. The problem of determining the right sense of a word (*w.r.t* some standard lexicon), given the context, is called Word Sense Disambiguation (WSD) [Yarowsky, 1992], [Kilgarriff, 1998], [Agirre, 1996], [McRoy, 1992], [Nancy, 1999]. The meaning that a particular word takes, depends on the meanings that the surronding words take, given that the surrounding words may also be ambiguous. An intuition is that even though many words in the text may have multiple meanings, they all have a unique sense in the context such that these senses reinforce each other. Take the following sentence as an example : *I reached the bank using the boat.* The word *bank* has 10 meanings according to the WordNet [Fellbaum, 1998] - an electronic lexical database, while the word *boat* has 2 senses. If one carefully goes through these senses, one will find that the *river bank* sense of *bank* and the *watercraft* sense of *boat* are related to each other through the concept of *river*. This should help us disambiguate each of these individual words. The first question that arises is: how does one discover the relations between *boat* and *river* on the one hand and *bank* and *river* on the other hand? One answer lies in the WordNet [Fellbaum, 1998] which provides a hypergraph structure connecting all related senses of different words using a variety of relational edges.

The next question that arises is : which are the words in a document that help in disambiguating each other ? Do **all words** in the document help in mutually disambiguating each other? In this paper, we make the assumptions that words that are closer to each other in the text, help in mutual disambiguation more than words that are far apart. More specifically - greater the distance between two words in the text, lesser is the possibility of them sharing a common context. For example, the words *bank* and *boat* are assumed to be associated with a *river* if they occur somewhere close together in the text (or in the same sentence). But if they occur "far away", they may or may not be referring to *river*. But in the context of distance between words in a document, what is "close" and what is "far"? In section 3 we describe the model we use to account for the distance between words in disambiguating them.

1.1.1 Semantic set

A piece of text has one or more subject matters of discussion. For example, the above sentence was about a river bank and not about a financial institution. The senses of the words that are related to each other in the WordNet hypergraph, should form a set - which we call the *semantic set* [Ganesh et al, 2002]. Different meanings of the different words in a text, taking into account the distances between the words, form many sets. But there is a set that is more dominant than the others and more representative of the text, and consisting of the true senses of the words in the text. Taking the above sentence as an example again, WordNet gives the following as 3 of the relevant *semantic sets*: **{ bank(financial institution) }, {bank(river bank),boat(watercraft), river}, {boat(dish for sauce)}** . The words in brackets indicate the sense of the corresponding word. Note that the word *river* is unambiguous - *i.e.* it has only one meaning according to the WordNet. A semantic set can be constructed for every word-sense with that word-sense as the *anchor* or *root word* and associating with it, all other word-senses in the text, that are related through the WordNet. In section 2, we formally define the idea of *semantic set* and in section 3 we describe a model for finding those semantic sets which are most representative of the theme or subject matter of the text. In the above example, there are two semantic sets *rooted* with different senses of the word *boat*, because there are two senses of the word *boat* in the WordNet. **(1)** {boat(watercraft), bank(river bank),river}, **(2)** {boat(dish for sauce)} . For the purpose of finding the "most representative" semantic set corresponding to each sense of every word, we assign scores to them (section 3). These scores are based on the textual distance as well as the semantic distance (or WordNet distance) between the component words of the *semantic set*. The scores of the *semantic sets* are assigned in a global manner - accoss all the senses of all the words in the text.

1.2 Keyword Extraction

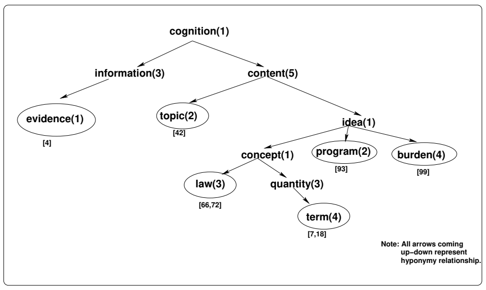
Keyphrases provide a brief summary of a document's contents. They have several applications like representing the contents of a document in a query engine, document similrity measurement etc. In the specific domain of keyphrases, there are two fundamentally different approaches: keyphrase assignment and keyphrase extraction. Keyphrase assignment seeks to select the phrases from a controlled vocabulary that best describe a document. Keyphrase extraction [Turney, 1999], the approach used here, does not use a controlled vocabulary, but instead chooses keyphrases from the text itself. **According to the model proposed in this paper**, the word-senses with relatively high scores are more representative of the text than other word-senses. The corresponding words can form the *keywords* for the text. In section 4 we discuss how these scores are used to extract keyphrases from a text.

2 Terminology

Below is a sample text of 100 words, from the Brown Corpus, with some nouns underlined. In the WordNet sub-graph of figure 1, the relationship between these nouns is shown. The words marked in ellipses are words that actually occur in the text. The number in the brackets, by the side of each word, is its WordNet sense number. The numbers mentioned in the square brackets are the textual positions. For example, the word *law* appears in textual positions 66 and 72. The arrows going up-down show the hyponymy relations. Thus, 2 hyponyms of sense number 1 of *cognition* are shown. In the same way, one can consider the *holonymy*, *meronymy*, *synonymy* and *antonymy* relations in the WordNet to capture all the nouns in a given piece of text. Consider the resultant WordNet sub-graph. Suppose that distances are measured over edges, with every edge of unit distance and the distances are additive. Consider all the words that occur in the graph, within a distance of 4 from the 2^{nd} sense of the word *program*. We call the set of word-senses, within a fixed distance from the chosen synset as the *semantic set* corresponding to that synset. Fig. 2(B) is an example. We want to find the correct senses of the words in a text T. Let W be a window in T having n nouns, $(w_1, w_2, w_3, ...w_n)$. For every w_i its s_i senses are $\sigma_{i_1},......\sigma_{i_{s_i}}$. Let P_i be the position of w_i in the text.

The Fulton_County_Grand_Jury said Friday an investigation of Atlanta 's recent primary_election produced no *evidence* that any irregularities took_place. The jury further said in *term* end presentments that the City_Executive_Committee which had over-all charge of the election deserves the praise and thanks of the City_of_Atlanta for the manner in which the election was conducted The September-October *term* jury had been charged by Fulton Superior_Court_Judge_Durwood_Pye to investigate reports of possible irregularities in the hard-fought primary which was won by Mayor-nominate_Ivan_Allen_Jr ..... It recommended that Fulton legislators act to have these laws studied and revised to the end of modernizing and improving them. The grand_jury commented on a number of other *topics* among them the Atlanta and Fulton_County purchasing_departments which it said are well operated and follow generally accepted practices which inure to the best interest of both governments However the jury said it believes these two offices should be combined to achieve greater efficiency and reduce the cost of administration ... Implementation of Georgia 's automobile title *law* was also recommended by the outgoing jury It urged that the next Legislature provide enabling funds and re-set the effective date so_that an orderly implementation of the *law* may be effected. ... This is one of the major items in the Fulton_County general assistance *program* the jury said but the State_Welfare_Department has seen_fit to distribute these funds through the welfare departments of all the counties in the state with the exception of Fulton_County which receives none of_this money The jurors said they realize a proportionate distribution of these funds might disable this *program* in our less populous counties. The jurors said Failure to do this will continue to place a disproportionate *burden* on Fulton taxpayers.

(A)



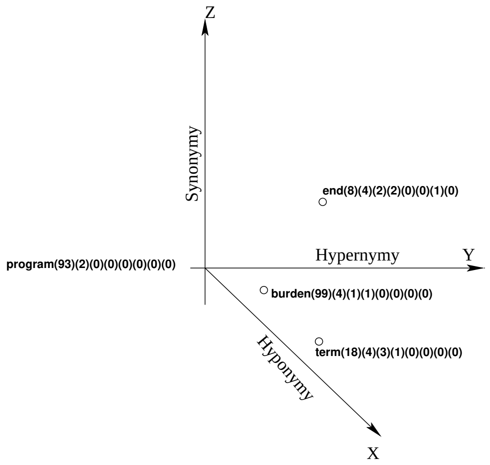
(B)

Figure 1: (A)Sample text from SemCor, br-a01 with the word *program* (word number 93) in consideration and (B) an extract of the WordNet graph, corresponding to the nouns underlined in figure (A)

1. **Semantic Graph:** Let G be that minimal sub-graph of the WordNet, which includes all the noun-senses σ_{i_k}, $1 \leq k \leq s_i$ and $1 \leq i \leq n$, from T. We call G, the *Semantic Graph* for the text T. Let σ_{i_p} and σ_{j_q} be two noun-senses in the sub-graph G. Consider the shortest path from σ_{i_p} to σ_{j_q}.

program<80,2,<0,0,0,0,0,0>>, portion<95,1,<0,3,1,0,1,0>>, policy<58,2,<0,2,0,0,0,0>>,
term<7,4,<0,3,1,0,0,0>>, topic<42,2,<0,1,2,0,0,0>>, law<66,1,<0,3,1,0,0,0>>,
end<8,3,<0,4,0,1,0,0>>, term<18,4,<0,3,1,0,0,0>>, end<8,4,<0,1,0,1,0,1>>,
practice<45,5,<0,1,3,0,0,0>>, manner<14,3,<0,4,1,0,0,0>>, law<66,3,<0,2,1,0,0,0>>,
end<39,3,<0,4,0,1,0,0>>, end<39,4,<0,1,0,1,0,1>>, law<72,1,<0,3,1,0,0,0>>,
burden<99,4,<0,1,1,0,0,0>>, city<59,1,<0,2,3,0,0,0>>, law<72,3,<0,2,1,0,0,0>>,
city<31,1,<0,2,3,0,0,0>>, city<56,1,<0,2,3,0,0,0>>, evidence<4,1,<0,2,3,0,0,0>>

(A)



(B)

Figure 2: (A) Semantic set corresponding to sense number 2 of the word *program* (B) (B)3-D Graph showing the relative positions of three words with respect the the word *program*

Let $\eta_1(\sigma_{i_p}, \sigma_{j_q})$, $\eta_2(\sigma_{i_p}, \sigma_{j_q})$, $\eta_3(\sigma_{i_p}, \sigma_{j_q})$, $\eta_4(\sigma_{i_p}, \sigma_{j_q})$, $\eta_5(\sigma_{i_p}, \sigma_{j_q})$ and $\eta_6(\sigma_{i_p}, \sigma_{j_{q_6}})$ respectively be the number of *hyponymy*, *hypernymy*, *meronymy*, *holonymy*, *synonymy* and *antonymy* arcs on this path.

2. **Semantic vector:** The *semantic vector* between two noun senses σ_{i_p} and σ_{j_q} in the graph G is the sequence $(\eta_1(\sigma_{i_p}, \sigma_{j_q})), ...(\eta_6(\sigma_{i_p}, \sigma_{j_q}))$ where the $\eta_i(\sigma_i, \sigma_{j_q})$'s are as given in previous definition. We denote the semantic vector by $N(\sigma_{i_p}, \sigma_{j_q})$.

3. **Semantic distance** The concept of *semantic distance* has been explored in [Alexander et al, 2001]. Broadly, two concepts of *semantic distance* have been mentioned there. They are *semantic similarity* and *semantic relatedness*. In this paper, we talk of semantic relatedness, as explored in [Graeme et al , 1998]. But the measures of *semantic distance* that we adopt are little variants of what has been proposed by [Graeme et al , 1998]. The first measure of the *semantic distance* of a noun-sense σ_{j_q} from σ_{i_p} in G corresponds to the minimum number of arcs that must be traversed in order to reach σ_{j_q} from σ_{i_p}. From the fact that *hypernymy*, *hyponymy* and *meronymy*, *holonymy* are complementary, and that *synonymy* and *antonymy* are *symmetric*, it follows that the semantic distance is *commutative*. The second measure of *semantic distance* will be given in section **??**.

4. **Semantic form** Recall the definition that P_i is the position of w_i in the text. The expression, $w_j < P_j, q, < N(\sigma_{i_p}, \sigma_{j_q}) >>$ is called the *semantic form* for σ_{j_q} *with respect to* σ_{i_p}. We will denote it by $F(\sigma_{i_p}, \sigma_{j_q})$.

5. **Semantic set** Consider every noun-sense σ_{j_q} in G, within a maximum semantic distance of R from σ_{i_p}. The collection of all the semantic forms $F(\sigma_{i_p}, \sigma_{j_q})$ is called the *semantic set* S_{i_p} for σ_{i_p}, with radius R. σ_{i_p} is called the *reference noun-sense* for S_{i_p}. A semantic set S_{i_p} is of the form $S_{i_p} = F(\sigma_{i_p}, \sigma_{i_p}), F(\sigma_{i_p}, \sigma_{i_{1_{p1}}})....F(\sigma_{i_p}, \sigma_{i_{k_{p_k}}})$ where k is the length of the semantic set. For

word w_i we have s_i semantic sets S_{i_1}, S_{i_2} , ... $S_{i_{s_i}}$. Also, for the word sense σ_{i_p} we define the position vector $\overline{P_{i_p}}$ as $\overline{P_{i_p}} =< P_{i_1}...P_{i_k} >$ and $\overline{N_{i_p}}$ as $\overline{N_{i_p}} =< N(\sigma_{i_p}, \sigma_{i_{1\,p_1}}), ...N(\sigma_{i_p}, \sigma_{i_{k\,p_k}}) >$

The wordsenses of the underlined nouns in the text in the figure 1, form a semantic graph, part of which has been depicted in figure 1. For instance consider the word *program* which has position number 93 in br-a01 and *burden* which has position number 99 in br-a01. In figure 1, it is shown that sense number 4 of *burden* and sense number 2 of *program* have the same hypernym - the sense number 1 of *idea*. Thus, the semantic distance between *program(2)* and *burden(4)* is 2. The semantic vector from *program(2)* to *burden(4)*, keeping *program(2)* as the reference word is $(\eta_1(\sigma_{93_2}, \sigma_{99_4}))$,.. $(\eta_6(\sigma_{93_2}, \sigma_{99_4})) =< 0, 1, 1, 0, 0, 0 >$ The semantic form $F(\sigma_{93_2}, \sigma_{99_4})$ is given as *burden*$< 99, 4, < 0, 1, 1, 0, 0, 0 >>$. σ_{99_4} is within a semantic distance of 4 from σ_{93_2}. The collection of all $F(\sigma_{93_2}, \sigma_{j_q}), 1 \leq q \leq s_j \forall$ words $\sigma_j, j \neq i$ in the text T such that, σ_{j_q} is within a semantic distance 4 from σ_{93_2} is called the semantic set for σ_{93_2}, $S(\sigma_{93_2})$. This semantic set is given in figure 2(A).

3 The Approach

The problem of finding the appropriate sense for w_i can be transformed to the problem of choosing the corresponding appropriate semantic set for w_i. This means we intend to find a measure function $M(S_{i_p}) = m_{i_p}$ such that $\text{argmax}_{1 \leq p \leq s_i} M(S_{i_p})$ gives the correct sense for the word w_i. The idea is that, a word-sense in the text indicates the presence of other word-senses in the piece of text in such a way that **semantically close word senses should also appear textually close**. Therefore, a word-sense in the text is affected by another word-sense in the text in two ways. First is that, *as the semantic distance between them increases, the influence should decrease.* Secondly, *as the textual distance between them increases, the influence should decrease.* Intuitively, the first factor plays a predominant role in determining the sense of the word under consideration. This follows from the fact that **slight variation in textual position of a word-sense does not influence the sense of the passage as such. But a slight variation in semantic distance considerably alters the sense of the passage.** Based on these two observations, we state the following hypothesis:

Hypothesis

The measure $M(S_{i_p})$ is of the form $M(\overline{P_{i_p}}, \overline{N_{i_p}})$. The contribution of each word σ_{i_j} in the semantic set to the score $M(S_{i_p})$ decreases exponentially with the semantic distance from w_i and decreases inversely with its textual distance from w_i.

We consider two semantic distances: the *Manhattan distance*, $H(\sigma_{i_p}, \sigma_{j_q})$ and the eucledian distance $H'(\sigma_{i_p}, \sigma_{j_q})$ as given in equation 1.

$$H(\sigma_{i_p}, \sigma_{j_q}) = \Sigma_{m=1}^{6} \eta_m(\sigma_i, \sigma_{j_q}), or, H'(\sigma_{i_p}, \sigma_{j_q}) = \sqrt{\Sigma_{m=1}^{6} (\eta_m(\sigma_i, \sigma_{j_q})^2)} \tag{1}$$

According to the hypothesis mentioned above, the expression for the measure function is as given in equation 2. Note that expressions containing $H(\sigma_{i_p}, \sigma_{j_q})$ could be replaced by $H'(\sigma_{i_p}, \sigma_{j_q})$

$$M(\overline{P_{i_p}}, \overline{N_{i_p}}) = \Sigma_{F(\sigma_{i_p}, \sigma j_q) \in S(\sigma_{i_p})} \frac{1}{|P_{j_q} - P_i|} \times e^{-H(\sigma_{i_p}, \sigma_{j_q})} \tag{2}$$

For a word w_i, the appropriate sense number is p and the second most appropriate sense number is $\overline{p}$ iff the conditions given in equation (3) are satisfied.

$$p = \underset{0 \leq j \leq s_i}{\operatorname{argmax}} M(S_{i_j}), \overline{p} = \underset{0 \leq j \leq s, j \neq k}{\operatorname{argmax}} M(S_{i_j}) \tag{3}$$

Mutual Reinforcement

We may note that **a word w which has a unique sense in WordNet, helps disambiguate other words related to it**. That is, if word w_j has only one WordNet sense, we would like to give special attention to this information, in all the sets that contain σ_{j_1}. For instance, if the p^{th} semantic set for w_i, i.e S_{i_p} has the word w_j, with w_j having only one sense in the WordNet, giving more weight to w_j, sense number 1, will add additional emphasis on the p^{th} sense of w_i. Moreover, we would like that this effect on σ_{i_p} be reflected on all the sets that contain σ_{i_p}.in turn. To ensure that this happens, we make the following changes to equation 2. Initially, we set the score for each semantic set to 1. Next, within the semantic sets for a word, we normalize the scores. Not that sets corresponding to unambiguous word senses (i.e word senses for the words having just one WordNet sense) will have a score of 1 initially. Then we find the new measure for each semantic set using equation 4:

$$I(\sigma_{i_p}, \sigma_{j_q}) = M(S_{j_q}) \times \frac{1}{|P_{j_q} - P_i|} \times e^{-H(\sigma_{i_p}, \sigma_{j_q})}, M(\overline{P_{i_p}}, \overline{N_{i_p}}) = \Sigma_{F(\sigma_{i_p}, \sigma j_q) \in S(\sigma_{i_p})} I(\sigma_{i+p}, \sigma_{j_q}) \tag{4}$$

After updating all set measures for w_i using equation 4, we normalize the measures for the sets corresponding to w_i using equation 5.

$$M(S_{i_p}) = M(\overline{P_{i_p}}, \overline{N_{i_p}}) = \frac{M(S_{i_p})}{\Sigma_{r=1}^{s_i} M(S_{i_r})} \tag{5}$$

Note that in the equation 4 we have scaled the entry for each term σ_{j_q} in the set S_{i_p}, by the measure $M(S_{j_q})$ for the corresponding set S_{j_q}. This means that, if in a particular iteration, sense number q of w_j is found to be more probable than that of the other senses of w_j, then it's contribution to the scores of other sets is more than the other senses of w_j.

4 Keyword Extraction

One advantage of using the method proposed in this paper is that we do not need any training corpus. Our method does not solely depend on the frequency of occurance of the words in the text as in purely

*INTIALISATION

1. Incrementally construct semantic sets S_{i_p}, $1 \leq p \leq i_s$, for each of the i_s WordNet senses of w_i, $1 \leq i \leq n$.

2. for all $1 \leq i \leq n$ do

   (a) for all $1 \leq p \leq s_i$ do

      i. $M(S_{i_p}) = \frac{1}{s_i}$ /* Note that we have combined 2 steps into 1; setting $M(S_{i_p})$ to 1 and then normalising */

(A)

1. do till the scores stabilise

   - **MUTUAL REINFORCEMENT**

   (a) for all $1 \leq i \leq n$ do

      i.

      ii. for all $1 \leq p \leq s_i$ do

         A. $H(\sigma_{i_p}, \sigma_{j_q}) = \Sigma_{m=1}^{6} \eta_m(\sigma_i, \sigma_{j_q})$ /*This could be replaced by the Eucledian measure.*/

         B. $M(\overline{P_{i_p}}, \overline{N_{i_p}}) = \Sigma_{F(\sigma_{ip}, \sigma j_q) \in S(\sigma_{ip})} M(S_{j_q}) \times \frac{1}{|P_{jq} - P_i|} \times e^{-H(\sigma_{ip}, \sigma_{jq})}$

   - **NORMALISATION**

   (b) for all $1 \leq i \leq n$ do

      i. for all $1 \leq p \leq s_i$ do

         A. $M(S_{i_p}) = M(\overline{P_{i_p}}, \overline{N_{i_p}}) = \frac{M(S_{ip})}{\Sigma_{r=1}^{s_i} M(S_{ir})}$

(B)

Figure 3: (A) The INITIALIZATION Pseudo-code for the method (B) The MUTUAL REINFORCEMENT Pseudo-code for the method

statistical methods. Rather, the frequency of occurance of the words is automatically captured in the construction of *semantic sets*. Semantic sets provide a more elegant way of capturing the relation between words occuring in a text in terms of their word-senses, rather than the statistical measures which rely on the occurance of certain words in the training corpus. A particular word may occur with different meanings (polysemy) or different words may occur with same meanings (synonymy). A method is proposed herein to capture this phenomena through the construction of semantic sets.

We will first discuss *keyword extraction*. Then we discuss *keyphrase extraction* as a natural consequence of *keyword extraction*. Our hypothesis is as follows "A word can be a keyword for a document if the sense in which it is being used is a predominating sense in the document".The word-sense corresponding to the semantic set with the highest overall measure is the most dominating sense for the whole documents, according to the model on which we build the word sense disambiguation algorithm. Therefore, to find the top "n" keywords for a document, all we need to do is find the "n" semantic sets with highest scores and pick the corresponding words. An automatic impliation is the we are stating that the "n" keywords are being used in the "n" particular senses which form the roots of the "n" highest measure semantic sets.

We will again work with the text of figure 1. Three of the highest scoring semantic sets that were constructed are given in figure 4(A). We can see quite clearly that most of the keywords, like *georgia, grand_jury, administration, state, primary, irregularity* are clustered together in the highest scoring synsets. The manually extracted keywords for this test were *primary, election, grand_jury, irregularity, registration, government, state.* We do not consider proper nouns like *Ivanr Allen Jr* in the semantic sets, for two reasons. One is that the WordNet does not contain entriees for all proper nouns. For example, it contains an entry for *Georgia* but no entry for *Fulton County Jury.* So there is presently no way of including such word-senses in the semantic sets. Secondly, proper nouns can be extracted as keywords using IR techniques such as the naive bayes [Witten et al, 1999]. In our experiments, we extracted all occurances of proper nouns as keywords. An important observation is that word frequencies are captured in an elegant way through the construction of *semantic sets.* In br-a01, the word *jury* occurs more than once. All the occurances of *jury*(WordNet sense 1) have been captured in the semantic set rooted at *government*(WordNet sense 1)

atlanta<43,1,<0>>,georgia<33,1,<1>>,law<66,1,<4>>,state<86,1,<2>>, law<72,1,<4>>,state<86,2,<3>>,
city<59,1,<2>>,city<31,1,<2>>, city<56,1,<2>>, georgia<63,1,<1>> =====> 0.0550942805644137

grand_jury<73,1,<0>>,jury<25,1,<0>>,juror<97,1,<2>>,jury<32,1,<0>>, jury<67,1,<0>>,legislature<68,1,<4>>,
jury<81,1,<0>>,administration<52,2,<3 >>, jury<6,1,<0>>,government<47,1,<4>>,jury<53,1,<0>>,jury<19,1,<0>>,
juror<90,1,<2>>,
size<30,1,<4>>,jury<48,1,<0>> =====> 1.50303676103437

government<47,1,<0>>,interest<46,7,<4>>,jury<25,1,<3>>,investigation<1,1,<4>>, juror<97,1,<4>>,jury<32,1,<3>>,jury<67,1,<3>>,
interest<26,7,<4>> , administration<52,1,<2>>,legislature<68,1,<1>>,jury<81,1,<3>>,administration<52,2,<1>>,
jury<6,1,<3>>,
office<49,1,<3>>,office<49,2,<3>>,person nel<57,1,<4>>, state<86,1,<1>>,jury<53,1,<3>>,jury<19,1,<3>>,juror<90,1,<4>>,
personnel<54,1,<4>>,welfare<83,1,<4>>,department<84,1,<1>>,department<84,2,<2>>, jury<48,1,<3>>
=====> 1.58303039218139

(A)

program<93,1,<0,0,0,0,0,0>>, ... evidence<4,1,<0,2,2,1,0,0>> = 0.312, **0.023**
program<93,2,<0,0,0,0,0,0>>, ... portion<95,1,<0,3,1,0,1,0>> = 1.129, **0.088**
program<93,3,<0,0,0,0,0,0>>, ... election<35,2,<0,3,1,0,1,0>> = 0.144, **0.009**
program<93,4,<0,0,0,0,0,0>>, ... report<24,1,<0,1,1,0,0,0>> = 0.899, **0.611**
program<93,5,<0,0,0,0,0,0>>, ... title<65,1,<0,1,1,1,0,0>> = 0.186, **0.017**
program<93,6,<0,0,0,0,0,0>>, . distribution<91,3,<0,1,3,1,0,0>> = 1.107, **0.053**
program<93,7,<0,0,0,0,0,0>>, ... laws<38,1,<0,2,2,0,0,1>> = 0.629, **0.045**
program<93,8,<0,0,0,0,0,0>>, ... city<56,1,<0,1,3,0,0,1>> = 0.473, **0.022**

(B)

Figure 4: (A)3 of the semantic sets with the highest scores(B)Example of 8 semantic sets for the word *program*

We have been talking in terms of keywords rather than key-phrases because key-phrases are typically composed of keywords. In the above example, *primary election* was found to be a key-phrase. But our method gives both *primary* as well as *election* as keywords. We use co-occurance of keywords to form key-phrases. Thus, n keywords that occur together in the text are considered to form 1 key-phrase.

5 Experiments and results

5.1 Word Sense Disambiguation

Experiments for evaluating our procedure for word sense disambiguation were performed over nouns in Brown corpus and checked against SemCor for correctness. As an example case, consider the 93^{rd} noun, *program* in the text in figure 1. It is tagged with sense number 2 in SemCor. Figure 4(B) shows the 8 semantic sets for the word *program.* Using equations 1 and 2, we get the scores for the different sets as indicated by the bold number to the right of each set in the figure. *The scores stabilize after around 10 iterations.* We find highest score for the second set - thus indicating *sense number 2.* Thus, the algorithm correctly disambiguated the word *program.* The experiments were carried out on the first 100 nouns for each of 10 documents from the *Brown corpus.* 2 tests were done - (1) comparing the top ranked sense p and (2) comparing the 2 top ranked senses, p and $\overline{p}$ derived using equation 3. The results for 5 of them are tabulated below in table ??. The average *precision* obtained using the

Eucledian measure was $3-4\%$ lower than that obtained using the *Manhattan measure*. The algorithm mentioned in this paper, performed better than that of [Agirre, 1996] (precision = 66.4 %, Recall = 58.8%). The window size $|W| = n$, for all the above tests was chosen as 100. Changing it to 150 produced improvement by $5-7\%$ and an average of 6.3%.

Table 1: Results (only some are documented for brevity) for WSD on Brown Corpus

Text	Coverage (%)	Prec. with top sense(%)	Rcl with top sense(%)	Prec with top 2 senses(%)	Rcl with top 2 senses(%)
a01	99	70	69.3	83.8	83.0
a02	98	69	67.6	75.5	74.0
a12	95	65.0	61.8	74.7	71.0
a13	95	62.0	58.90	74.2	72.0

5.2 Keyphrase Extraction

We carried out an empirical evaluation of our method using documents from the New Zealand Digital Library. Our goal was to assess the overall effectiveness of semantic sets. We measured keyphrase quality by counting the number of matches between our output and the keyphrases that were originally chosen by the document's author. Kea [Witten et al, 1999], a state-of-the-art keyphrase extractor has been evaluated using the Computer Science Technical Reports (CSTR) collection of the NZDL. From the 46,000 documents in this corpus, we chose 80 where the author had supplied keyphrases. During each run of the experiment, we performed 2 different measures. First we split all the keyphrases given by the author, into individual words and considred these words as keywords. The other comparison we did was to compare the number of keyphrases found out by our procedure against those given by the author. We restricted the comparison to as many keywords as were generated by the author, in each case. On an average, the number of keywords expected per document were 18.77, while the number of keyphrases expected per document were 11.71. The results for a few of the texts have been documented in table 2.

Table 2: Evaluating our procedure against author extracted keyphrases

	Author		Our algo	
	# kws	# kphs	# kws	# kphs
CSTR/10970.txt	5	14	2	10
CSTR/11152.txt	7	15	3	10
CSTR/18904.txt	7	15	2	11
CSTR/19013.txt	6	14	3	9

On the whole, across all documents, of the 1502 keywords extracted by our algorithm, 913 keywords were found to be correct. This gave a precision of 60.79%. Of the 937 keyphrases found, 319 were found to be correct. This gives a precision of 34.05% for keyphrases. When compared against the performance of the KEA [Witten et al, 1999] extractor on the same corpus, we find that our method performs much better. The precision of KEA ranges from 9.40% to 18.60%, depending on the number of keyphrases extracted.

References

[Ganesh et al, 2002] Ganesh Ramakrishnan and Pushpak Bhattacharyya Word Sense Disambiguation Using Semantic Sets Based on WordNet. In *Proceedings of Workshop on Use of Semantics in Information Retrieval and Filtering*, LREC 2002, Las Palmas, Canary Islands, Spain.

[Yarowsky, 1992] David Yarowsky, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora". In *Proceedings of the* 14^{th} *International Conference on Computational Linguistics (COLING-92), pages 454-460, Nantes, France, 1992.*

[Kilgarriff, 1998] Adam Kilgarriff. Gold Standard Data-sets for Evaluating Word Sense Disambiguation Programs. In *Computer Speech and Language 12 (4), Special Issue on Evaluation, 1998.*

[Agirre, 1996] Agirre, E. and Rigau, G. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96.*

[Resnik, 1997] Phil Resnik. Selectional preference and sense disambiguation. In *Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington D.C., USA, 1997.

[Nancy, 1999] Nancy Ide. Parallel Translations as Sense Discriminators. In *Proceedings of SIGLEX99, Washington D.C, USA, 1999.*

[Green, 1997] Green, S.J. Automatically Generating Hyper-text by Computing Semantic Similarity. *Ph.D. Thesis, University of Toronto, 1997.*

[Fellbaum, 1998] Fellbaum, Christiane, ed. WordNet: An Electronic Lexical Database. *MIT Press, May 1998.*

[McRoy, 1992] McRoy.S. Using multiple knowledge sources for Word Sense Disambiguation. In *Computational Linguistics 18.1 (1992), pages 1-30.*

[Alexander et al, 2001] Alexander Budanitsky, Graeme Hirst. Semantic Distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, June 2001.

[Graeme et al , 1998] Graeme Hirst, David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In *Christiane Fellbaum (editor), WordNet: An electronic lexical database*, Cambridge, MA: The MIT Press, 1998, 305-332.

[Turney, 1999] Turney, P. Learning to extract keyphrases from text. In *Journel of Information Retrieval*, 1999.

[Witten et al, 1999] Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. KEA: Practical automatic keyphrase extraction. In *Proc. DL '99*, pp. 254-256.

[Frank et al, 1998] Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.G. Domain-specific keyphrase extraction. In *Proc. Sixteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668-673.