Text Representation with WordNet Synsets using Soft Sense Disambiguation

Ganesh Ramakrishnanan
hare@cse.iitb.ac.in
Computer Sc. & Engg.
Indian Institute of Technology
Mumbai - 400076

Pushpak Bhattacharyya
pb@cse.iitb.ac.in
Computer Sc. & Engg.
Indian Institute of Technology
Mumbai - 400076

Abstract: Text information processing depends critically on the proper representation of texts. A common and naive way of representing a text is as a bag of its component words. This representation suffers primarily from two drawbacks, *viz., polysemy* and *synonymy* which arise because of the ambiguity of the words and the lack of information about the relations between the words. This paper presents a model for representing a text in terms of the *synsets* in the *WordNet*- the lexical knowledge base of English words along with the semantic relations. These synsets stand for concepts which correspond to the words of the text. In particular, a *soft sense disambiguation* approach has been proposed. The text representation so obtained is found to convey the key ideas that the texts deal with. WordNet relations with other words in the sentence are exploited to disambiguate the senses. This scheme has been evaluated using a *goodness measure* based the *information content* of the representation of the text. As an actual application, the problem of *text classification* has been taken up, and the results are encouraging.

Keywords: WordNet, Synonymy, Polysemy, Semantic Graph, Synset-Ranking, Document Vectors, Hubs and Authorities, Bayesian Belief Networks, Mutual Information, Text Classification

1 Introduction

Representation of texts is critical in text information processing tasks like retrieval, classification, clustering, summarization, question-answering *etc.* Some common text representation schemes are *bag of words* [Nigam et al1999] [Dumais et al1998] and *web pages as a set of links to other pages* [Page et al1998] [Kleinberg1998].

Words in documents have multiple meanings (*polysemy*), or several words can have the same meaning (*synonymy*). For example, as a noun, the word *bank* is polysemous. In the sense of river bank, the words *bank*, *camber*, *river side etc*, are close synonyms. The problem of correct sense determination (with respect to a standard lexicon) in a context is called *Word Sense Disambiguation (WSD)* [Yarowsky1992] [Agirre and Rigau1996]

[Ganesh and Pushpak2001]. The meaning of a word depends on the meanings of the surrounding words which in turn may be ambiguous. An important observation is that *meaning emerges* through mutual sense reinforcement of possibly ambiguous words. For example, let us take the sentence *I reached the bank using the boat*. *Bank* has 10 senses according to the WordNet [Fellbaum1998], while *boat* has 2 senses. The *river bank* sense of *bank* and the *watercraft* sense of *boat* are related through the *river* concept. This should help us disambiguate these individual words. The first question, however, is *how does one discover the relationships between* boat *and* river *on one hand and that between* bank *and* river *on the other*? Approaches like *Latent Semantic Indexing (LSI)* [Dumais et al1998] rely on *frequent co-occurrence* of words as an approximate disambiguator. However, there is no way of detecting word similarity if they do not co-occur frequently enough; nor is it possible to detect polysemous usage of words if some senses of a word are rarely used.

Thus keeping in view the fact that *WSD* is a major hurdle to be crossed, we ask *if there be a method for capturing the essential information in a text, without requiring any training data*. In particular, we propose a *soft sense disambiguation* approach. One possibility of finding the answer lies in using the WordNet hyper graph structure connecting related senses of words through different semantic relations.

In the following section (section 2), we discuss the WordNet which is the foundation for our work and also review the related work. Section 3 deals with the problem of generating text representations. Section 5 describes various schemes for ranking synsets that are used for the representation. Evaluation and results are presented in section 7. Section 8 concludes the paper.

2 WordNet and related work

The English WordNet [Fellbaum1998] is an online lexical reference system whose design is inspired by current psycho-linguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets or *synsets*, each representing one underlying lexical concept. Noun synsets are related to each other through *hypernymy* (generalization), *hyponymy* (specialization), *holonymy* (whole of) and *meronymy* (part of) relations. Of these, (*hypernymy*, *hyponymy*) and (*meronymy*, *holonymy*) are complementary pairs.

The verb and adjective synsets are very sparsely connected with each other. No relation is available between noun and verb synsets. However, 4500 adjective synsets are related to noun synsets with *pertainyms* (pertaining to) and *attra* (attributed with) relations.

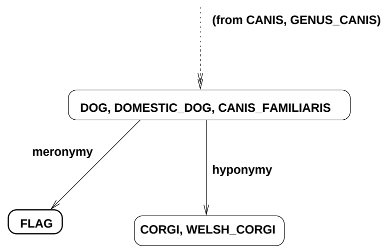

Figure 1: Illustration of the WordNet relations.

Figure 1 shows that the synset {*dog, domestic_dog, canis_familiaris*} has a hyponymy link to {*corgi, welshcorgi*} and meronymy link to {*flag*} ("a conspicuously marked or shaped tail").

We use adjectives whose synsets are related to noun synsets with *pertainymy* (pertaining to) or to noun words through *attribute* (attribute of) relations. For example, the adjective synset {*Biblical, scriptural*} pertains to the noun synset {*bible, Good Book, Holy Scripture, Holy Writing, Scripture, Word of God*} while the adjective synset {*beautiful*} is an attribute of the noun word *beauty*.

Attempts have been made to use the WordNet in information retrieval. [Scott et al1998] discusses the use of hypernyms to represent text in terms of the synsets of its constituent words. But it uses word-sense disambiguation as a separate module to first determine the *correct* synsets of the words in the text. [Agirre and Rigau1996] attempts to disambiguate the words in a text using the idea of *conceptual density* in the WordNet. These methods rely on making *hard* decisions on word senses - retaining only one sense per word. Hard decisions may be a requisite for tasks like Machine Translation, where one needs to commit to a word meaning before finding an equivalent word in the target language. This paper is different in its approach in that it uses the WordNet to simply identify *concepts* or *synsets* that are most relevant to the text and ranks them according *how they help relate words in the text to each other*. No *hard* decision on the senses of words are made, since word sense disambiguation is recognized as an implicit task rather than an end in itself. This is especially important since the state-of-the-art WSD systems do not perform very well on general texts. As the example in this paper illustrates, sense disambiguation is implicitly done by identifying densely connected regions in the semantic graph for the text. By this we hope to avoid making mistakes arising from *hard* decisions on word senses.

3 Generating text representation

In the discussions that follow, we restrict our attention to the noun part of the WordNet. We also consider the mapping (or association) of adjectives to the noun synsets by *pertainyms* and *attribute of* relations. Also we consider only *hypernymy* and *holonymy* relations and their inverses, *viz.*, *hyponymy* and *meronymy* respectively.

3.1 Notation

Let T be a text containing words $(\omega_1, \omega_2, ..\omega_n)$ (nouns and adjectives) associated with synsets $\sigma_1, \sigma_2...\sigma_m$ in the WordNet. We view the WordNet as a directed graph, $G = (V, E)$. V is the set of vertices and E is the set of directed arcs. $V = (\sigma_1, \sigma_2....\sigma_N)$. $(\sigma_i, \sigma_j) \in E$ *iff* either σ_i is a *hyponym* of σ_j or σ_i is a *meronym* of σ_j. The synset nodes which have no edges incident upon them will be hereafter called *root synsets*. There exist 14 such root synsets, in the noun WordNet, each of which have no *hypernyms*.

We call as *basis synsets*, those synsets to which words in the text belong. The graph of synsets that can be reached by traversing the hypernymy and holonymy links starting from all the *basis synsets* and upto the root will be called the *semantic graph* (G_T) for the text T. Suppose G_T has p nodes. Let $GRAPH_{p\times p}$ be the adjacency matrix for the DAG G_T. $GRAPH(i, j) = 1$ *iff* there is an edge from σ_i to σ_j in G_T. Else $GRAPH(i, j) = 0$.

3.2 An example

The word *bank* is ambiguous. Figures 2 and 3 show two different definitions of the word *bank* based on the WordNet glosses. Each word has, by its side, a number in parentheses showing the number of synsets in the WordNet in which it appears. Thus *bank* has 10 senses, *body* has 9 senses and so on. Word or term based representation will detect some similarity between T_1 and T_2, whereas, we would like them to have little or no similarity. *Also, two texts dealing with similar or related topics, but having no words in common should have a similar set of highly ranked features.*

Definition 1 Bank(10) is a geological(1) formation(5) on the sides(12) of a water(7) body(9) especially land(10) with a slope(2). It is a natural(14) formation(5). It could be a beach(1) on the side(12) of the sea(3) or the ocean(2) or a descent(6) on the sides(12) of a river(1) or lake(3).

Figure 2: Document T_1 : The definition of *bank* as a *river-bank*

Definition 2 Bank(10) is an institution(4). It lends money(3) to business(1) establishments(7) and plays(17) an important(1) role(4) in commerce(3).

Figure 3: Document T_2 : The definition of *bank* as a *financial_institution*

Figure 4, shows the semantic graph G_{b1} that contains the different synsets for the words in figure 2(T_1). Nodes correspond to synsets, and arcs correspond to *hyponymy* and *meronymy* relations. For every node in the graph, we have a node-id. Similarly a graph G_{b2} can be identified for the text T_2. The graph G_{b1} consists of 284 nodes and 324 edges while G_{b2} consists of 213 nodes and 228 edges.

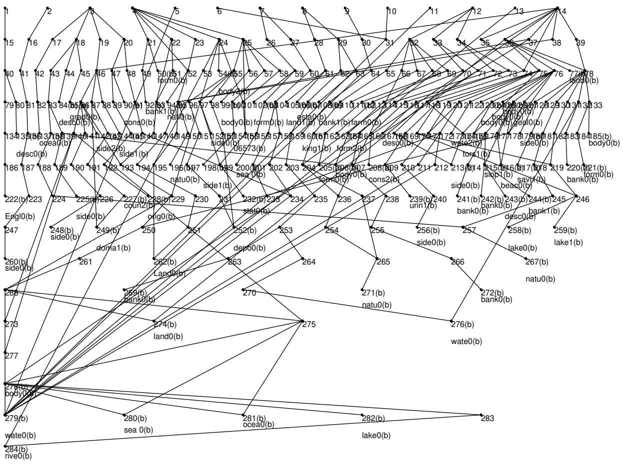

Figure 4: Semantic graph G_{b1} for the text T_1 (figure 2)

Our whole *work is centered around collecting topic-specific synsets arising out of the matter in the text. In particular, a soft decision is made with respect to the inclusion of a synset in this collection*. This approach, termed as *soft sense disambiguation*, is described in section 4.

4 Soft sense disambiguation

Word sense disambiguation is defined as the task of finding *the* sense of a word in a context. In this paper, we explore the idea that one should not commit to a particular sense of the word, but to a *set of its senses* which are not necessarily orthogonal or mutually exclusive. Very often the WordNet gives for a word multiple senses which are related and which *help connect* related words in the in the text amongst themselves, which we refer to as the *relevance of the sense* . Therefore, instead of picking a single sense, we rank the senses according to the degree of their relevance to the text. The following example illustrates this point.

In the table 3 for text T_2, the synset $\{$*depository_financial_institution, bank, banking_concern, banking_company*$\}$ with node-id 193 has the gloss from the WordNet as "a financial institution that accepts deposits and channels the money into lending activities". The synset $\{$*institution, organization*$\}$ can be reached by traversing the hypernymy link from this synset. Thus, the words, *institution* and *bank* in the text are related through the synset for *bank* in the *financial institution* sense.

Another synset with node-id 206 is $\{$*bank*$\}$ which has the WordNet gloss, "a supply or stock held in reserve for future use (especially in emergencies)". This sense may not correspond to the meaning of *bank* as used in the definition. But it is relevant to the definition, since it deals with *money* or *possession*. In fact, the synset $\{$*possession*$\}$ is reached by traversing 2 hypernymy links from the synset $\{$*money*$\}$ corresponding to the wealth sense of money and 2 hypernymy links from this synset for $\{$*bank*$\}$. In other words, the sense in question of $\{$*bank*$\}$ relates the words *bank* and *money* to each other.

In section 5, we present 3 algorithms to rank the *basis synsets* in the text and also rank all synsets that can be reached through these *basis synsets*. The ranking of synsets is subsequently evaluated in section 7. Since the similarity between words is implicitly captured in the rank of their connecting synsets, we represent text as a vector of the ranked synsets and subject this representation to two evaluations- an information theoretic measure of information content and a representative task of text classification.

5 Synset ranking algorithms

The following three algorithms are used to rank the synsets in the semantic graph representing the text.

1. Hubs and Authorities
2. Page Ranking
3. Bayesian Inferencing

5.1 Hubs and Authorities algorithm for synset ranking

This algorithm is motivated by the page-ranking algorithm [Kleinberg1998] on the World Wide Web. We call *authorities*, those synsets that provide significant and useful information on the topic of the text. For instance $\{$*bank, cant, camber*$\}$, $\{$*body of water, water*$\}$, $\{$*slope, incline, side*$\}$ are some of the potential authorities for T_1. Similarly, $\{$*depository institution, bank*$\}$, $\{$*business*$\}$, $\{$*financial activity, commercial activity*$\}$ are some of the potential authorities for T_2. We call as *hubs*, those synsets that provide lots of useful links to relevant content synsets (topic *authorities*$\}$. $\{$*geological formation, formation*$\}$, $\{$*natural object, artifact*$\}$ are few of the potential hubs for T_1 while, $\{$*social group*$\}$ and $\{$*institution, establishment*$\}$ are some potential hubs for T_2.

The task at hand is to computationally determine hubs and authorities for a particular text through analysis of its semantic graph. We notice two mutually recursive facts: *That hubs point to lots of authorities and that authorities are pointed to by lots of hubs. Together they tend to form a bipartite graph*.

We use an iterative algorithm to converge on a mutually reinforcing set of hubs and authorities. Let p be the number of synsets in the semantic graph G_T. We maintain for each synset $\sigma_i \in G_T$ an authority score: $A(i)$ and a hub score $H(i)$, where H and A are vectors of size p. We initialize as in equation 5.1.

$$H(i) = A(i) = 1, \forall i. \tag{1}$$

We maintain normalized scores $||H||_2 = ||A||_2 = 1$, where $||H||_2$ refers to the L_2 norm of H. Authorities are pointed to by many good hubs.

$$A(i) = \sum_{j|GRAPH(j,i)=1} H(j) \tag{2}$$

that is, $A = GRAPH^T.H$. Also hubs point to many good authorities:

$$H(j) = \sum_{j|GRAPH(j,i)=1} A(i) \tag{3}$$

that is, $H = GRAPH.A$. A and H are computed by initializing with equation 5.1 and iterating over equations 5.1 and 5.1. It is a standard result of linear algebra [Golub and Loan1989] [Kleinberg1998] that the iterations asymptotically converge to a fix-point. Specifically, it can be established that the vector A converges to the principal eigenvector of $GRAPH^T.GRAPH$ and that the hub vector H converges to the principal eigenvector of $GRAPH.GRAPH^T$. We maintain two rankings of the synsets *viz.* one according to A and the other according to H.

The distinction between hubs and authorities, however, is rather artificial. For a given text, a synset can act as both a hub, and an authority. In conformity with intuition, we found that hubs generally tend to be located towards the top of the hyponymy-meronymy hierarchy and authorities, towards the bottom of this hierarchy. Hubs are generalized concepts. An alternative method of synset ranking is to not to attempt to capture the distinction between hubs and authorities. This is because, it may be very difficult to point out in a semantic graph which are the hubs and which are the authorities. For instance, a text may contain a word belonging to one of the *root* synsets in the WordNet like {*entity*}, {*psychological feature*} or {*abstraction*}. Next we try to rank synsets just by their authority scores.

5.2 Page Ranking algorithm for ranking synsets

This approach is motivated by Google's page-ranking algorithm [Page et al1998] (called PAGERANKING algorithm). We note that the in-degree alone is not a sufficient indicator of the authority. Because these incoming links may be one of the many outgoing links from the source synsets. Therefore, we take into account, the out-degree of the synset nodes.

Consider a text T and the semantic graph G_T for T. Let R be a vector such that $R(i)$ is the authority rank for the i^{th} synset in G_T. Let $GRAPH$ be the adjacency matrix of G_T. Let N_i be the out-degree of the synset σ_i. We define a new matrix A of size $i \times i$ as $A(i,j) = \frac{1}{N_i}$ *iff* $GRAPH(i,j) = 1$ else $A(i,j) = 0$. A represents the graph as per the intuition given above. A synset, σ_i, 'gives' an equal fraction of its authority to all the synsets it points to. The following equations show the initialization and the iterative algorithm for computing R.

$$R_{old}(i) = 1,\ if\ \sigma_i\ is\ a\ basis\ synset\ else\ 0 \tag{4}$$

$$R_{new}(i) = \frac{1}{||R_{new}||_2} \sum_{j|GRAPH(j,i)=1} \frac{R_{old}(j)}{N_j} \tag{5}$$

that is, $R_{new} = \frac{A^T \times R_{old}}{||A^T \times R_{old}||_2}$. One can view it as a process of synset-ranks 'flowing' from synsets to the synsets they point to. It can be proved that R converges to the principal eigenvector of A. Alternatively one can view synset-rank as modeling a 'random surfer' that starts on a random synset and then at each point randomly follows a link on the current synset. $R(i)$ models the probability that this random surfer will be on synset σ_i at any given time.

5.3 Bayesian inferencing for synset ranking

In this section, we explore the use of a kernel built using WordNet, to map from word-space to synset-space. We build a Bayesian Belief Network(BBN) [Heckerman1995] from the semantic graph for this purpose. We first explain in brief, what a Bayesian Belief Network(BBN) is. A BBN for a set of random variables $X = X_1, X_2, ...X_m$ consists of a network of m nodes, each node representing one random variable with directed arcs between the nodes. The network structure encodes a set of conditional independence assertions about variables in X and a set of local probability distributions associated with each variable. Missing arcs between nodes encode independencies, conditioned on their parents, such that

$$p(X_1 = x_1, X_2 = x_2, ...X_m = x_m) = \Pi_{i=1}^{m} p(X_i = x_i | PA_i = pa_i) \tag{6}$$

where PA_i is a vector of 'parent nodes' for node X_i. The local probability distribution for a node, given its parents, is the term $p(X_i = x_i | PA_i = pa_i)$.

With every synset $\sigma_i, (1 \leq i \leq p)$, in a semantic graph, we associate a random variable X_{σ_i}, whose value corresponds to the *relevance* of σ_i to the text *i.e.*, to its *synset-score*. In addition, we associate a random variable X_{ω_j} with each word $\omega_j, 1 \leq j \leq n$ that occurs in the text. The BBN consists of the random variables, $X_{\sigma_1}, X_{\sigma_2} ....X_{\sigma_p}, X_{\omega_1}, ...X_{\omega_n}$. Directed edges between nodes are added as follows. A directed edge is drawn from X_{σ_i} to X_{σ_j} if there is a directed edge from σ_i to σ_j in the semantic graph. An arc is drawn from X_{σ_i} to X_{ω_j} if ω_j appears in σ_i.

We assume a conditional gaussian distribution for each node, with parameters mean $\mu = 1$ and variance $\sigma = 0.25$. The gaussian distribution for a random variable X_i with continuous parents PA_i is given as

$$f(x_i | pa_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x_i - (W^T . pa_i)^2}{2\sigma^2}} \tag{7}$$

where W is a vector of weights on the links from parents PA_i of X_i to X_i, denoting their relative importance. W^T is the transpose of W. We set all the weights to 1 since we have no apriori knowledge of which of the parents of X_i is a more probable "cause" for X_i.

To rank the synsets, the X_{ω_i} nodes are clamped at 1. Next, bayesian inferencing is carried out to find the expected value of each of the internal nodes X_{σ_j}. Inferencing was done using the BayesNet toolbox (version as of 14^{th} November, 2002) [Murphy2001]. The higher the expected value, the higher is the rank of the synset.

6 Observation on T_1 and T_2

The results of applying the synset ranking algorithms to G_{b1} and G_{b2} and locating the top 20 highly synsets in each case have been shown in figures 1, 2, 3 and 4. The results from the page ranking algorithm are found to be slightly better.

It may be noted that the word *bank* has been disambiguated in each case when either the PAGERANKING or the Bayesian inferencing algorithms are used. By disambiguation, we mean the assignment of the highest score to the correct synset. Also, many other synsets related to the correct synset for *bank* have got high ranks. In fact, the overall ranking for the synsets showed that all the words in the two texts T_1 and T_2 were disambiguated in the *soft* sense described in 4.

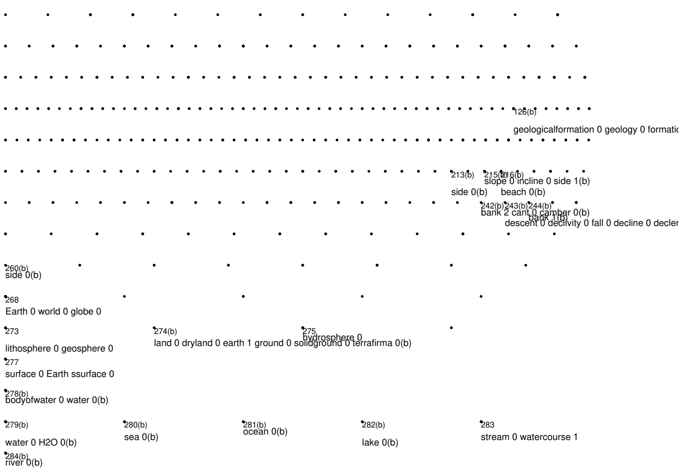

Figure 5: Semantic graph for text T_1 showing synsets ranked top by the PAGERANKING algorithm

Figure 5 shows some of the highly ranked synsets obtained by ranking them using the PAGERANKING algorithm. Note that the edges in the graph 4 have been removed in 5 for better visibility of the synsets. The reader may also note that highly ranked synsets are cluttered together in the graph - indicating that relevant synsets reinforce each other.

Hub Scores	Authority Scores
{universe, cosmos}	{water, H2O }
{collection, aggregation, assemblage}	{ocean}
{galaxy, extragalacticnebula}	{sea}
{group, grouping}	{lake}
{phenomenon}	{stream, watercourse}
{object, physicalobject}	{bodyofwater, water}
{geologicalformation, geology, formation }	{location}
planet	{unit, buildingblock}
{celestialbody, heavenlybody}	{object, physicalobject}
{universe, existence, creation, world,}	{artifact, artefact}
{naturalobject}	{land, earth, ground, terrafirma }
{entity}	{naturalobject}
{system}	{hydrosphere}
{slope, incline, side}	{surface, Earth surface}
{naturalelevation, elevation}	{causalagent, cause, causalagency}

Table 1: The top-ranked 15 synsets for the definition of *bank* as in figure 2, obtained by the first 2 algorithms.

Pagerank Scores	Bayesian Regression
{sea}	{sea}
{ocean}	{ocean}
{water, H2O }	{water, H2O}
{river}	{river}
{lake}	{lake}
{stream, watercourse}	{stream, watercourse}
{bodyofwater, water}	{bodyofwater, water}
{slope, incline, side}	{land, earth, ground, terrafirma}
{bank, cant, camber }	{natural}
{descent, declivity, fall, decline, downslope }	{bank, cant, camber}
{bank}	{bank}
{beach}	{descent, declivity, fall, decline, downslope}
{Earth, world, globe}	
{side}	{side}
{geologicalformation, geology, formation }	{surface, Earth ssurface}

Table 2: The top-ranked 15 synsets for the definition of *bank* as in figure 2, obtained by the last 2 algorithms.

Hub Scores	Authority Scores
{activity}	{aim, object, objective, target}
{act, humanaction, humanactivity}	{executivebranch}
{organization, organisation}	{accumulation}
{diversion, recreation}	{administration, organisation }
{institution, establishment }	{sector}
{commercialenterprise, business }	{governmentdepartment}
{socialgroup}	{bankingindustry}
{commerce, commercialism, mercantilism }	{idea, thought}
{group, grouping}	{possession}
{finance}	{act, humanaction, humanactivity}
{importance}	{group, grouping}
{artifact, artefact}	{event}
{transaction, dealing, dealings}	{pointofreference, reference}
{action}	{psychologicalfeature}
{establishment}	{evasion}

Table 3: The top-ranked 15 synsets for the definition of *bank* as in figure 3 obtained by the first 2 algorithms.

Pagerank Scores	Bayesian Regression
{DepartmentofCommerce}	{executivedepartment}
{executivedepartment}	{DepartmentofCommerce}
{federaldepartment}	{executivebranch}
{executivebranch}	{federaldepartment}
{depositoryfinancialinstitution, bank}	{governmentdepartment}
{governmentdepartment}	{department, section}
{business}	{business_sector}
{finance}	{branch, subdivision, arm}
{money}	{division}
{branch, subdivision, arm}	{branch, subdivision, arm}
{department, section}	{depositoryfinancialinstitution, bank}
{businessactivity, commercialactivity}	{money}
{bank}	{finance}
{division}	{bank}
{significance}	{business, businesssector}

Table 4: The top-ranked 15 synsets for the definition of *bank* as in figure 3 obtained by the last 2 algorithms.

7 Evaluation

We represent a text as a vector of synset scores determined by the algorithms above. The features in the text vector are the synsets and the feature values are the synset scores found using one of the algorithms described in this paper. Two ways of evaluating this representation and the results thereof are presented.

7.1 Based on mutual information

This section is based on the work of Rong Jin [Jin et al2001] which shows empirically that the information content of the document vector representation, has a direct positive bearing on performance in information retrieval tasks. We show that the information content we obtain for the synset vector representation of documents is higher than that of the tfidf representation of documents. We now proceed to define the information content measure.

Let $d_1, d_2, ..., d_n$ be the document vectors in a particular feature space. Let *M* be the document-feature matrix. Each number M_{ij} in the matrix *M* represents the weight of the j^{th} feature in the i^{th} document. Let *D* be the document-document matrix defined as $D = M^T.M$. The eigenvectors u_i of *D* form an orthonormal basis for the column space of D and M. Let *C* be a random vector having distribution $P(C = u_i) = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$. C is called the *document content vector* [Jin et al2001] suggests the use of $I(C, D)$, the mutual information between the random vector C and the random document-document matrix D, as a measure of the 'goodness' of the feature representation used for the documents. Interested readers can refer to [Jin et al2001] for details.

20,000 documents from the 20-NewsGroups data-set were considered for this experiment.

This data-set has 20 classes with 1000 documents per class. Part of speech tagging was done using the *qtagger*. The $I(C, D)$ values for different feature representations of M have been sbown in 5.

The result of applying the algorithm on the document vector representations described in this paper is summarized in the table 5. The document vector obtained using the PAGERANKING algorithm for synset ranking gives the maximum information content. The vector obtained using the Bayesian inferencing approach does nearly as good as the PAGERANKING algorithm. It is noticed that the hub-scores are not very informative about the document content.

Document representation	**I(C,D)**
TF	**1.3563**
TFIDF	**1.9873**
Ranking using the PAGERANKING method	**3.4372**
Ranking with hub-scores	**1.9633**
Ranking with authority-scores	**2.6386**
Ranking with BayesNet approach	**3.4251**
Raking using conceptual vectors	**2.7261**

Table 5: Information content measures for various feature representations for 20-NewsGroups data-set

7.2 Evaluation with classification

Classification experiments were performed on the same *NewsGroups* data-set using the $k-$ *nearest neighbor* ($k - nn$) classifier was used. The results are presented in figure 6. It is found that the scoring performed using PAGERANKING algorithm for synset-ranking (section 5.2) performs the best followed closely by the scoring using Bayesian inferencing.

The above graph shows that, *with even a very small training sample (only 2-10 documents per class) the performance for classification obtained using synset representation, is much better than that obtained using term representation*. The reason is that for a small number of training documents per class, the vocabulary per class is very small for the term representation, whereas for the synset vector representation of documents, the vocabulary of the class is expanded, in a weighted manner, using WordNet. As a result, with synset representation for a document, we can detect similarity of that document with documents from the same class, even if they have very few words in common between them. With a term representation for a document, on the other hand, a document will be given a low similarity measure with documents from the same class if they have only few common words. We expect that as the size of the training sample increases, the margin between performances obtained using the 2 representation schemes will narrow down, since the per-class vocabulary will also increase. Whereas, the synset-based similarity detection method will not be that much sensitive to the occurrence of matching words. We can observe that tendency in figure 6; as the number of per class training documents is increased, the margin between performances of the two representation schemes starts decreasing.

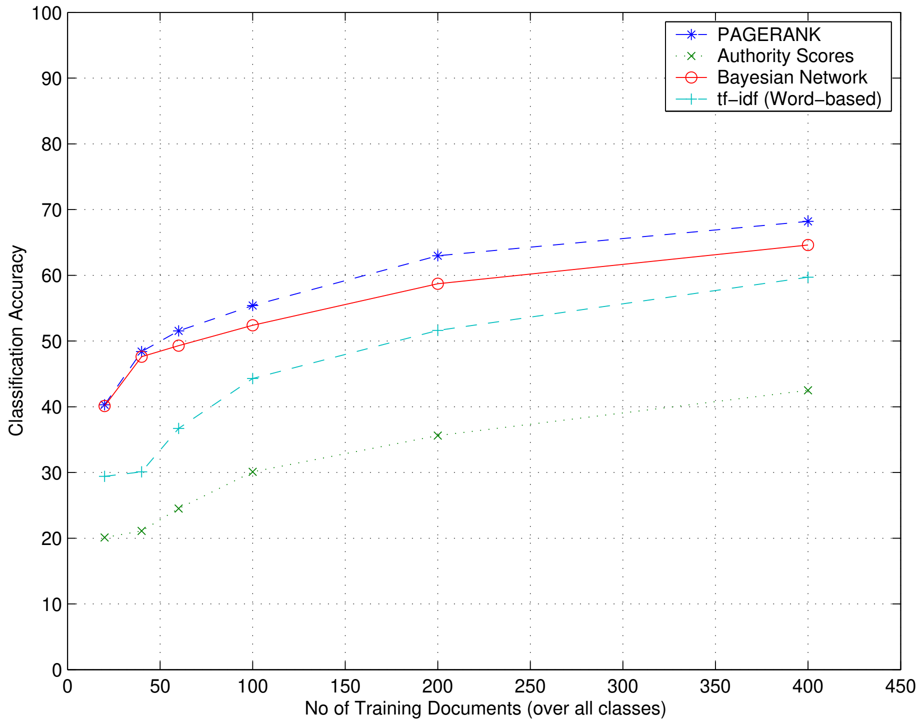

Figure 6: Classification accuracy for two-fold cross validation with different number of documents (positive examples) for training

8 Conclusions

In this paper a novel idea for representing text for information retrieval has been presented. In particular, a *soft sense disambiguation* paradigmn was described. The WordNet hypergraph was exploited extensively. The goodness of the representation has been calculated using mutual information. The vectors have been subjected to a representative information retrieval task, *viz.*, text classification. Results show that the feature scores generated using the PAGERANKING algorithm for synset-ranking serve the purpose the best, followed closely by the bayesian inferencing approach. The conclusion is that WordNet does helps relate the words in a document and in the emergence of meaning through mutual reinforcement of related words. This method of ranking synsets for a text can find use in many other applications like clustering, question answering and summarization, some of which are ongoing. Future work consists in assigning weights top the edges of the semantic graph and incorporating verbs.

References

[Fellbaum1998] Fellbaum Christiane, ed. the WordNet: An Electronic Lexical Database. *MIT Press*, Map 1998.

[Yarowsky1992] 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.

[Agirre and Rigau1996] Agirre, E. and Rigau, G. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*.

[Ganesh and Pushpak2001] Ganesh Ramakrishnan and Pushpak Bhattacharyya. Word Sense Disambiguation Using Semantic Sets Based on the WordNet. *International Conference for Language Resource Evaluation: Special Workshop on Using Semantics for Information Retrieval and Filtering*, Canary Islands, June, 2001.

[Scott et al1998] Scott, Sam and Stan Matwin. Text classification using the WordNet hypernyms. In *Proceedings of the COLING/ACL Workshop on Usage of the WordNet in Natural Language Processing Systems*, Montreal, 1998.

[Kleinberg1998] Jon M. Kleinberg. Authoritative Sources in a Hyper-linked Environment. In *Journal of the ACM (1998).*

[Page et al1998] Page, L., Brin, S., Motwani, R., and Terry Winograd, T. The Pagerank citation ranking: Bringing order to the web. *Technical report, Stanford*, (Santa Barbara, CA 93106, January 1998)

[Heckerman1995] David Heckerman. A Tutorial on Learning with Bayesian Networks. In *Technical Report MSR-TR-95-06*, Microsoft Research, March, 1995.

[Murphy2001] Kevin Murphy. The Bayes Net Toolbox for Matlab. In *Computing Science and Statistics*, vol 33, 2001.

[Jin et al2001] Rong Jin, Christos Falusos, Alex G. Hauptmann. Meta-scoring: automatically evaluating term weighting schemes in IR without precision-recall. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, New Orleans, Louisiana, United States.

[Golub and Loan1989] Golub,G., Van Loan, C. F. *Matrix Computations*, Johns Hopkins University Press, 1989, Baltimore, Md.

[Dumais et al1998] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing*, New York: ACM, 281-285.

[Nigam et al1999] Kamal Nigam, John Lafferty and Andrew McCallum. Using Maximum Entropy for Text Classification. *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999.