A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness

George Tsatsaronis and **Vicky Panagiotopoulou**
Department of Informatics
Athens University of Economics and Business,
76, Patision Str., Athens, Greece
`gbt@aueb.gr, vpanagiotopoulou@gmail.com`

Abstract

Generalized Vector Space Models (GVSM) extend the standard Vector Space Model (VSM) by embedding additional types of information, besides terms, in the representation of documents. An interesting type of information that can be used in such models is semantic information from word thesauri like WordNet. Previous attempts to construct GVSM reported contradicting results. The most challenging problem is to incorporate the semantic information in a theoretically sound and rigorous manner and to modify the standard interpretation of the VSM. In this paper we present a new GVSM model that exploits WordNet's semantic information. The model is based on a new measure of semantic relatedness between terms. Experimental study conducted in three TREC collections reveals that semantic information can boost text retrieval performance with the use of the proposed GVSM.

1 Introduction

The use of semantic information into text retrieval or text classification has been controversial. For example in Mavroeidis et al. (2005) it was shown that a GVSM using WordNet (Fellbaum, 1998) senses and their hypernyms, improves text classification performance, especially for small training sets. In contrast, Sanderson (1994) reported that even 90% accurate WSD cannot guarantee retrieval improvement, though their experimental methodology was based only on randomly generated pseudowords of varying sizes. Similarly, Voorhees (1993) reported a drop in retrieval performance when the retrieval model was based on WSD information. On the contrary, the construction of a sense-based retrieval model by Stokoe

et al. (2003) improved performance, while several years before, Krovetz and Croft (1992) had already pointed out that resolving word senses can improve searches requiring high levels of recall.

In this work, we argue that the incorporation of semantic information into a GVSM retrieval model can improve performance by considering the semantic relatedness between the query and document terms. The proposed model extends the traditional VSM with term to term relatedness measured with the use of WordNet. The success of the method lies in three important factors, which also constitute the points of our contribution: 1) a new measure for computing semantic relatedness between terms which takes into account relation weights, and senses' depth; 2) a new GVSM retrieval model, which incorporates the aforementioned semantic relatedness measure; 3) exploitation of all the semantic information a thesaurus can offer, including semantic relations crossing parts of speech (POS). Experimental evaluation in three TREC collections shows that the proposed model can improve in certain cases the performance of the standard TF-IDF VSM. The rest of the paper is organized as follows: Section 2 presents preliminary concepts, regarding VSM and GVSM. Section 3 presents the term semantic relatedness measure and the proposed GVSM. Section 4 analyzes the experimental results, and Section 5 concludes and gives pointers to future work.

2 Background

2.1 Vector Space Model

The VSM has been a standard model of representing documents in information retrieval for almost three decades (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999). Let D be a document collection and Q the set of queries representing users' information needs. Let also t_i symbol-

ize term i used to index the documents in the collection, with $i = 1, .., n$. The VSM assumes that for each term t_i there exists a vector $\vec{t_i}$ in the vector space that represents it. It then considers the set of all term vectors $\{\vec{t_i}\}$ to be the generating set of the vector space, thus the space basis. If each d_k,(for $k = 1, .., p$) denotes a document of the collection, then there exists a linear combination of the term vectors $\{\vec{t_i}\}$ which represents each d_k in the vector space. Similarly, any query q can be modelled as a vector $\vec{q}$ that is a linear combination of the term vectors.

In the standard VSM, the term vectors are considered pairwise orthogonal, meaning that they are linearly independent. But this assumption is unrealistic, since it enforces lack of relatedness between any pair of terms, whereas the terms in a language often relate to each other. Provided that the orthogonality assumption holds, the similarity between a document vector $\vec{d_k}$ and a query vector $\vec{q}$ in the VSM can be expressed by the cosine measure given in equation 1.

$$cos(\vec{d_k}, \vec{q}) = \frac{\sum_{j=1}^{n} a_{kj} q_j}{\sqrt{\sum_{i=1}^{n} a_{ki}^2 \sum_{j=1}^{n} q_j^2}} \quad (1)$$

where a_{kj}, q_j are real numbers standing for the weights of term j in the document d_k and the query q respectively. A standard baseline retrieval strategy is to rank the documents according to their cosine similarity to the query.

2.2 Generalized Vector Space Model

Wong et al. (1987) presented an analysis of the problems that the pairwise orthogonality assumption of the VSM creates. They were the first to address these problems by expanding the VSM. They introduced term to term correlations, which deprecated the pairwise orthogonality assumption, but they kept the assumption that the term vectors are linearly independent[1], creating the first GVSM model. More specifically, they considered a new space, where each term vector $\vec{t_i}$ was expressed as a linear combination of 2^n vectors $\vec{m_r}$, $r = 1..2^n$. The similarity measure between a document and a query then became as shown in equation 2, where $\vec{t_i}$ and $\vec{t_j}$ are now term vectors in a 2^n dimensional vector space, $\vec{d_k}, \vec{q}$ are the document and the query

vectors, respectively, as before, a'_{ki}, q'_j are the new weights, and $\acute{n}$ the new space dimensions.

$$cos(\vec{d_k}, \vec{q}) = \frac{\sum_{j=1}^{\acute{n}} \sum_{i=1}^{\acute{n}} a'_{ki} q'_j \vec{t_i} \vec{t_j}}{\sqrt{\sum_{i=1}^{\acute{n}} a'^{2}_{ki} \sum_{j=1}^{\acute{n}} q'^{2}_j}} \quad (2)$$

From equation 2 it follows that the term vectors $\vec{t_i}$ and $\vec{t_j}$ need not be known, as long as the correlations between terms t_i and t_j are known. If one assumes pairwise orthogonality, the similarity measure is reduced to that of equation 1.

2.3 Semantic Information and GVSM

Since the introduction of the first GVSM model, there are at least two basic directions for embedding term to term relatedness, other than exact keyword matching, into a retrieval model: (a) compute semantic correlations between terms, or (b) compute frequency co-occurrence statistics from large corpora. In this paper we focus on the first direction. In the past, the effect of WSD information in text retrieval was studied (Krovetz and Croft, 1992; Sanderson, 1994), with the results revealing that under circumstances, senses information may improve IR. More specifically, Krovetz and Croft (1992) performed a series of three experiments in two document collections, CACM and TIMES. The results of their experiments showed that word senses provide a clear distinction between relevant and nonrelevant documents, rejecting the null hypothesis that the meaning of a word is not related to judgments of relevance. Also, they reached the conclusion that words being worth of disambiguation are either the words with uniform distribution of senses, or the words that in the query have a different sense from the most popular one. Sanderson (1994) studied the influence of disambiguation in IR with the use of pseudowords and he concluded that sense ambiguity is problematic for IR only in the cases of retrieving from short queries. Furthermore, his findings regarding the WSD used were that such a WSD system would help IR if it could perform with very high accuracy, although his experiments were conducted in the Reuters collection, where standard queries with corresponding relevant documents (qrels) are not provided.

Since then, several recent approaches have incorporated semantic information in VSM. Mavroeidis et al. (2005) created a GVSM kernel based on the use of noun senses, and their hypernyms from WordNet. They experimentally

---

[1] It is known from Linear Algebra that if every pair of vectors in a set of vectors is orthogonal, then this set of vectors is linearly independent, but not the inverse.

showed that this can improve text categorization. Stokoe et al. (Stokoe et al., 2003) reported an improvement in retrieval performance using a fully sense-based system. Our approach differs from the aforementioned ones in that it expands the VSM model using the semantic information of a word thesaurus to interpret the orthogonality of terms and to measure semantic relatedness, instead of directly replacing terms with senses, or adding senses to the model.

3 A GVSM Model based on Semantic Relatedness of Terms

Synonymy (many words per sense) and polysemy (many senses per word) are two fundamental problems in text retrieval. Synonymy is related with recall, while polysemy with precision. One standard method to tackle synonymy is the expansion of the query terms with their synonyms. This increases recall, but it can reduce precision dramatically. Both polysemy and synonymy can be captured on the GVSM model in the computation of the inner product between $\vec{t_i}$ and $\vec{t_j}$ in equation 2, as will be explained below.

3.1 Semantic Relatedness

In our model, we measure semantic relatedness using WordNet. It considers the path length, captured by *compactness* (SCM), and the path depth, captured by *semantic path elaboration* (SPE), which are defined in the following. The two measures are combined to for *semantic relatedness* (SR) beetween two terms. SR, presented in definition 3, is the basic module of the proposed GVSM model. The adopted method of building semantic networks and measuring semantic relatedness from a word thesaurus is explained in the next subsection.

Definition 1 *Given a word thesaurus O, a weighting scheme for the edges that assigns a weight $e \in (0, 1)$ for each edge, a pair of senses $S = (s_1, s_2)$, and a path of length l connecting the two senses, the semantic compactness of S ($SCM(S,O)$) is defined as $\prod_{i=1}^{l} e_i$, where $e_1, e_2, ..., e_l$ are the path's edges. If $s_1 = s_2$ $SCM(S,O) = 1$. If there is no path between s_1 and s_2 $SCM(S,O) = 0$.*

Note that *compactness* considers the path length and has values in the set [0, 1]. Higher *compactness* between senses declares higher semantic relatedness and larger weight are assigned to stronger edge types. The intuition behind the assumption of edges' weighting is the fact that some edges provide stronger semantic connections than others. In the next subsection we propose a candidate method of computing weights. The *compactness* of two senses s_1 and s_2, can take different values for all the different paths that connect the two senses. All these paths are examined, as explained later, and the path with the maximum weight is eventually selected (definition 3). Another parameter that affects term relatedness is the depth of the sense nodes comprising the path. A standard means of measuring depth in a word thesaurus is the hypernym/hyponym hierarchical relation for the noun and adjective POS and hypernym/troponym for the verb POS. A path with shallow sense nodes is more general compared to a path with deep nodes. This parameter of semantic relatedness between terms is captured by the measure of *semantic path elaboration* introduced in the following definition.

Definition 2 *Given a word thesaurus O and a pair of senses $S = (s_1, s_2)$, where $s_1, s_2 \in O$ and $s1 \neq s2$, and a path between the two senses of length l, the semantic path elaboration of the path ($SPE(S,O)$) is defined as $\prod_{i=1}^{l} \frac{2 d_i d_{i+1}}{d_i + d_{i+1}} \cdot \frac{1}{d_{max}}$, where d_i is the depth of sense s_i according to O, and d_{max} the maximum depth of O. If $s_1 = s_2$, and $d = d_1 = d_2$, $SPE(S,O) = \frac{d}{d_{max}}$. If there is no path from s_1 to s_2, $SPE(S,O) = 0$.*

Essentially, SPE is the harmonic mean of the two depths normalized to the maximum thesaurus depth. The harmonic mean offers a lower upper bound than the average of depths and we think is a more realistic estimation of the path's depth. SCM and SPE capture the two most important parameters of measuring semantic relatedness between terms (Budanitsky and Hirst, 2006), namely path length and senses depth in the used thesaurus. We combine these two measures naturally towards defining the *Semantic Relatedness* between two terms.

Definition 3 *Given a word thesaurus O, a pair of terms $T = (t_1, t_2)$, and all pairs of senses $S = (s_{1i}, s_{2j})$, where s_{1i}, s_{2j} senses of t_1, t_2 respectively. The semantic relatedness of T ($SR(T,S,O)$) is defined as $max\{SCM(S,O) \cdot SPE(S,O)\}$. SR between two terms t_i, t_j where $t_i \equiv t_j \equiv t$ and $t \notin O$ is defined as 1. If $t_i \in O$ but $t_j \notin O$, or $t_i \notin O$ but $t_j \in O$, SR is defined as 0.*
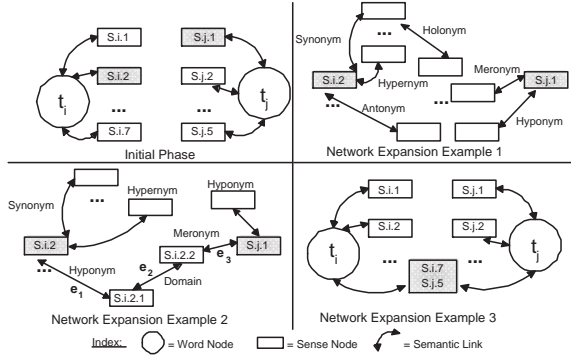
Figure 1: Computation of semantic relatedness.

3.2 Semantic Networks from Word Thesauri

In order to construct a semantic network for a pair of terms t_1 and t_2 and a combination of their respective senses, i.e., s_1 and s_2, we adopted the network construction method that we introduced in (Tsatsaronis et al., 2007). This method was preferred against other related methods, like the one introduced in (Mihalcea et al., 2004), since it embeds all the available semantic information existing in WordNet, even edges that cross POS, thus offering a richer semantic representation. According to the adopted semantic network construction model, each semantic edge type is given a different weight. The intuition behind edge types' weighting is that certain types provide stronger semantic connections than others. The frequency of occurrence of the different edge types in Wordnet 2.0, is used to define the edge types' weights (e.g. 0.57 for hypernym/hyponym edges, 0.14 for nominalization edges etc.).

Figure 1 shows the construction of a semantic network for two terms t_i and t_j. Let the highlighted senses $S.i.2$ and $S.j.1$ be a pair of senses of t_i and t_j respectively. All the semantic links of the highlighted senses, as found in WordNet, are added as shown in example 1 of figure 1. The process is repeated recursively until at least one path between $S.i.2$ and $S.j.1$ is found. It might be the case that there is no path from $S.i.2$ to $S.j.1$. In that case $SR((t_i, t_j), (S.i.2, S.j.1), O) = 0$. Suppose that a path is that of example 2, where e_1, e_2, e_3 are the respective edge weights, d_1 is the depth of $S.i.2$, d_2 the depth of $S.i.2.1$, d_3 the depth of $S.i.2.2$ and d_4 the depth of $S.j.1$, and d_{max} the maximum thesaurus depth. For reasons of simplicity, let $e_1 = e_2 = e_3 = 0.5$, and $d_1 = 3$. Naturally, $d_2 = 4$, and let $d_3 = d_4 = d_2 = 4$. Finally, let $d_{max} = 14$, which is the case for Word-

Net 2.0. Then, $SR((t_i, t_j), (S.i.2, S.j.1), O) = 0.5^3 \cdot 0.4615 \cdot 0.5^2 = 0.01442$. Example 3 of figure 2 illustrates another possibility where $S.i.7$ and $S.j.5$ is another examined pair of senses for t_i and t_j respectively. In this case, the two senses coincide, and $SR((t_i, t_j), (S.i.7, S.j.5), O) = 1 \cdot \frac{d}{14}$, where d the depth of the sense. When two senses coincide, $SCM = 1$, as mentioned in definition 1, a secondary criterion must be levied to distinguish the relatedness of senses that match. This criterion in SR is SPE, which assumes that a sense is more specific as we traverse WordNet graph downwards. In the specified example, $SCM = 1$, but $SPE = \frac{d}{14}$. This will give a final value to SR that will be less than 1. This constitutes an intrinsic property of SR, which is expressed by SPE. The rationale behind the computation of SPE stems from the fact that word senses in WordNet are organized into synonym sets, named *synsets*. Moreover, synsets belong to hierarchies (i.e., noun hierarchies developed by the hypernym/hyponym relations). Thus, in case two words map into the same synset (i.e., their senses belong to the same synset), the computation of their semantic relatedness must additionally take into account the depth of that synset in WordNet.

3.3 Computing Maximum Semantic Relatedness

In the expansion of the VSM model we need to weigh the inner product between any two term vectors with their semantic relatedness. It is obvious that given a word thesaurus, there can be more than one semantic paths that link two senses. In these cases, we decide to use the path that maximizes the semantic relatedness (the product of SCM and SPE). This computation can be done according to the following algorithm, which is a modification of Dijkstra's algorithm for finding the shortest path between two nodes in a weighted directed graph. The proof of the algorithm's correctness follows with theorem 1.

Theorem 1 *Given a word thesaurus O, a weighting function $w : E \rightarrow (0, 1)$, where a higher value declares a stronger edge, and a pair of senses $S(s_s, s_f)$ declaring source (s_s) and destination (s_f) vertices, then the $SCM(S, O) \cdot SPE(S, O)$ is maximized for the path returned by Algorithm 1, by using the weighting scheme $e_{ij} = w_{ij} \cdot \frac{2 \cdot d_i \cdot d_j}{d_{max} \cdot (d_i + d_j)}$, where e_{ij} the new weight of the edge connecting senses s_i and s_j, and w_{ij} the initial*

Algorithm 1 MaxSR(G,u,v,w)

Require: A directed weighted graph G, two nodes u, v and a weighting scheme $w : E \rightarrow (0..1)$.

Ensure: The path from u to v with the maximum product of the edges weights.

Initialize-Single-Source(G,u)

1: **for all** vertices $v \in V[G]$ **do**
2: $\quad d[v] = -\infty$
3: $\quad \pi[v] = NULL$
4: **end for**
5: $d[u] = 1$

Relax(u,v,w)

6: **if** $d[v] < d[u] \cdot w(u,v)$ **then**
7: $\quad d[v] = d[u] \cdot w(u,v)$
8: $\quad \pi[v] = u$
9: **end if**

Maximum-Relatedness(G,u,v,w)

10: Initialize-Single-Source(G,u)
11: $S = \emptyset$
12: $Q = V[G]$
13: **while** $v \in Q$ **do**
14: $\quad s = $ Extract from Q the vertex with max d
15: $\quad S = S \cup s$
16: $\quad$ **for all** vertices $k \in$ Adjacency List of s **do**
17: $\quad\quad$ Relax(s,k,w)
18: $\quad$ **end for**
19: **end while**
20: return the path following all the ancestors π of v back to u

weight assigned by weighting function w.

Proof 1 *For the proof of this theorem we follow the course of thinking of the proof of theorem 25.10 in (Cormen et al., 1990). We shall show that for each vertex $s_f \in V$, $d[s_f]$ is the maximum product of edges' weight through the selected path, starting from s_s, at the time when s_f is inserted into S. From now on, the notation $\delta(s_s, s_f)$ will represent this product. Path p connects a vertex in S, namely s_s, to a vertex in $V - S$, namely s_f. Consider the first vertex s_y along p such that $s_y \in V - S$ and let s_x be y's predecessor. Now, path p can be decomposed as $s_s \rightarrow s_x \rightarrow s_y \rightarrow s_f$. We claim that $d[s_y] = \delta(s_s, s_y)$ when s_f is inserted into S. Observe that $s_x \in S$. Then, because s_f is chosen as the first vertex for which $d[s_f] \neq \delta(s_s, s_f)$ when it is inserted into S, we had $d[s_x] = \delta(s_s, s_x)$ when s_x was inserted into S.*

We can now obtain a contradiction to the

above to prove the theorem. Because s_y occurs before s_f on the path from s_s to s_f and all edge weights are nonnegative[2] and in $(0,1)$ we have $\delta(s_s, s_y) \geq \delta(s_s, s_f)$, and thus $d[s_y] = \delta(s_s, s_y) \geq \delta(s_s, s_f) \geq d[s_f]$. But both s_y and s_f were in $V - S$ when s_f was chosen, so we have $d[s_f] \geq d[s_y]$. Thus, $d[s_y] = \delta(s_s, s_y) = \delta(s_s, s_f) = d[s_f]$. Consequently, $d[s_f] = \delta(s_s, s_f)$ which contradicts our choice of s_f. We conclude that at the time each vertex s_f is inserted into S, $d[s_f] = \delta(s_s, s_f)$.

Next, to prove that the returned maximum product is the $SCM(S,O) \cdot SPE(S,O)$, let the path between s_s and s_f with the maximum edge weight product have k edges. Then, Algorithm 1 returns the maximum $\prod_{i=1}^{k} e_{i(i+1)} =$

$$w_{s2} \cdot \frac{2 \cdot d_s \cdot d_2}{d_{max} \cdot (d_s + d_2)} \cdot w_{23} \cdot \frac{2 \cdot d_2 \cdot d_3}{d_{max} \cdot (d_2 + d_3)} \cdot ... \cdot w_{kf} \cdot$$
$$\frac{2 \cdot d_k \cdot d_f}{d_{max} \cdot (d_k + d_f)} = \prod_{i=1}^{k} w_{i(i+1)} \cdot \prod_{i=1}^{k} \frac{2 d_i d_{i+1}}{d_i + d_{i+1}} \cdot$$
$$\frac{1}{d_{max}} = SCM(S,O) \cdot SPE(S,O).$$

3.4 Word Sense Disambiguation

The reader will have noticed that our model computes the SR between two terms t_i, t_j, based on the pair of senses s_i, s_j of the two terms respectively, which maximizes the product $SCM \cdot SPE$. Alternatively, a WSD algorithm could have disambiguated the two terms, given the text fragments where the two terms occurred. Though interesting, this prospect is neither addressed, nor examined in this work. Still, it is in our next plans and part of our future work to embed in our model some of the interesting WSD approaches, like knowledge-based (Sinha and Mihalcea, 2007; Brody et al., 2006), corpus-based (Mihalcea and Csomai, 2005; McCarthy et al., 2004), or combinations with very high accuracy (Montoyo et al., 2005).

3.5 The GVSM Model

In equation 2, which captures the document-query similarity in the GVSM model, the orthogonality between terms t_i and t_j is expressed by the inner product of the respective term vectors $\vec{t_i}\vec{t_j}$. Recall that $\vec{t_i}$ and $\vec{t_j}$ are in reality unknown. We estimate their inner product by equation 3, where s_i and s_j are the senses of terms t_i and t_j respectively, maximizing $SCM \cdot SPE$.

$$\vec{t_i}\vec{t_j} = SR((t_i, t_j), (s_i, s_j), O) \qquad (3)$$

Since in our model we assume that each term can be semantically related with any other term, and

---
[2]The sign of the algorithm is not considered at this step.

$SR((t_i, t_j), O) = SR((t_j, t_i), O)$, the new space is of $\frac{n \cdot (n-1)}{2}$ dimensions. In this space, each dimension stands for a distinct pair of terms. Given a document vector $\vec{d_k}$ in the VSM TF-IDF space, we define the value in the (i, j) dimension of the new document vector space as $d_k(t_i, t_j) = (TF - IDF(t_i, d_k) + TF - IDF(t_j, d_k)) \cdot \vec{t_i}\vec{t_j}$. We add the TF-IDF values because any product-based value results to zero, unless both terms are present in the document. The dimensions $q(t_i, t_j)$ of the query, are computed similarly. A GVSM model aims at being able to retrieve documents that not necessarily contain exact matches of the query terms, and this is its great advantage. This new space leads to a new GVSM model, which is a natural extension of the standard VSM. The cosine similarity between a document d_k and a query q now becomes:

$$cos(\vec{d_k}, \vec{q}) = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} d_k(t_i, t_j) \cdot q(t_i, t_j)}{\sqrt{\sum_{i=1}^{n} \sum_{j=i}^{n} d_k(t_i, t_j)^2} \cdot \sqrt{\sum_{i=1}^{n} \sum_{j=i}^{n} q(t_i, t_j)^2}}$$

(4)

where n is the dimension of the VSM TF-IDF space.

4 Experimental Evaluation

The experimental evaluation in this work is two-fold. First, we test the performance of the semantic relatedness measure (SR) for a pair of words in three benchmark data sets, namely the Rubenstein and Goodenough 65 word pairs (Rubenstein and Goodenough, 1965)(R&G), the Miller and Charles 30 word pairs (Miller and Charles, 1991)(M&C), and the 353 similarity data set (Finkelstein et al., 2002). Second, we evaluate the performance of the proposed GVSM in three TREC collections (TREC 1, 4 and 6).

4.1 Evaluation of the Semantic Relatedness Measure

For the evaluation of the proposed semantic relatedness measure between two terms we experimented in three widely used data sets in which human subjects have provided scores of relatedness for each pair. A kind of "gold standard" ranking of related word pairs (i.e., from the most related words to the most irrelevant) has thus been created, against which computer programs can test their ability on measuring semantic relatedness between words. We compared our measure against ten known measures of semantic relatedness: (HS) Hirst and St-Onge (1998), (JC) Jiang and Conrath (1997), (LC) Leacock et al. (1998), (L) Lin (1998), (R) Resnik (1995), (JS) Jarmasz and Szpakowicz

(2003), (GM) Gabrilovich and Markovitch (2007), (F) Finkelstein et al. (2002), (HR)) and (SP) Strube and Ponzetto (2006). In Table 1 the results of SR and the ten compared measures are shown. The reported numbers are the Spearman correlation of the measures' rankings with the gold standard (human judgements).

The correlations for the three data sets show that SR performs better than any other measure of semantic relatedness, besides the case of (HR) in the M&C data set. It surpasses HR though in the R&G and the 353-C data set. The latter contains the word pairs of the M&C data set. To visualize the performance of our measure in a more comprehensible manner, Figure 2 presents for all pairs in the R&G data set, and with increasing order of relatedness values based on human judgements, the respective values of these pairs that SR produces. A closer look on Figure 2 reveals that the values produced by SR (right figure) follow a pattern similar to that of the human ratings (left figure). Note that the x-axis in both charts begins from the least related pair of terms, according to humans, and goes up to the most related pair of terms. The y-axis in the left chart is the respective humans' rating for each pair of terms. The right figure shows SR for each pair. The reader can consult Budanitsky and Hirst (2006) to confirm that all the other measures of semantic relatedness we compare to, do not follow the same pattern as the human ratings, as closely as our measure of relatedness does (low y values for small x values and high y values for high x). The same pattern applies in the M&C and 353-C data sets.

4.2 Evaluation of the GVSM

For the evaluation of the proposed GVSM model, we have experimented with three TREC collections [3], namely TREC 1 (TIPSTER disks 1 and 2), TREC 4 (TIPSTER disks 2 and 3) and TREC 6 (TIPSTER disks 4 and 5). We selected those TREC collections in order to cover as many different thematic subjects as possible. For example, TREC 1 contains documents from the Wall Street Journal, Associated Press, Federal Register, and abstracts of U.S. department of energy. TREC 6 differs from TREC 1, since it has documents from Financial Times, Los Angeles Times and the Foreign Broadcast Information Service.

For each TREC, we executed the standard base-

---

[3] http://trec.nist.gov/

	HS	JC	LC	L	R	JS	GM	F	HR	SP	SR
R&G	0.745	0.709	0.785	0.77	0.748	0.842	0.816	N/A	0.817	0.56	**0.861**
M&C	0.653	0.805	0.748	0.767	0.737	0.832	0.723	N/A	0.904	0.49	**0.855**
353-C	N/A	N/A	0.34	N/A	0.35	0.55	0.75	0.56	0.552	0.48	**0.61**

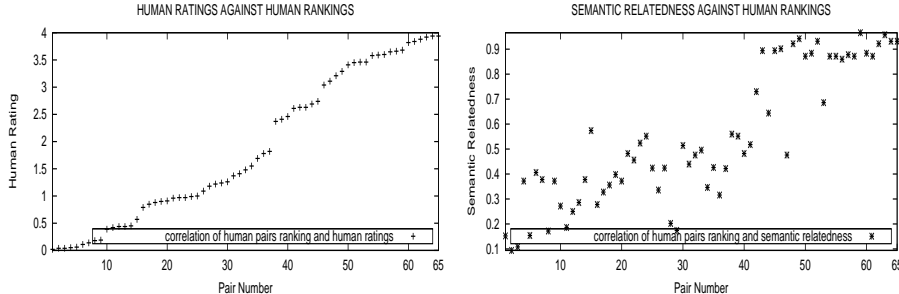Table 1: Correlations of semantic relatedness measures with human judgements.

Figure 2: Correlation between human ratings and SR in the R&G data set.

line TF-IDF VSM model for the first 20 topics of each collection. Limited resources prohibited us from executing experiments in the top 1000 documents. To minimize the execution time, we have indexed all the pairwise semantic relatedness values according to the SR measure, in a database, whose size reached 300GB. Thus, the execution of the SR itself is really fast, as all pairwise SR values between WordNet synsets are indexed. For TREC 1, we used topics $51 - 70$, for TREC 4 topics $201 - 220$ and for TREC 6 topics $301 - 320$. From the results of the VSM model, we kept the top-50 retrieved documents. In order to evaluate whether the proposed GVSM can aid the VSM performance, we executed the GVSM in the same retrieved documents. The interpolated precision-recall values in the 11-standard recall points for these executions are shown in figure 3 (left graphs), for both VSM and GVSM. In the right graphs of figure 3, the differences in interpolated precision for the same recall levels are depicted. For reasons of simplicity, we have excluded the recall values in the right graphs, above which, both systems had zero precision. Thus, for TREC 1 in the y-axis we have depicted the difference in the interpolated precision values (%) of the GVSM from the VSM, for the first 4 recall points. For TRECs 4 and 6 we have done the same for the first 9 and 8 recall points respectively.

As shown in figure 3, the proposed GVSM may improve the performance of the TFIDF VSM up to 1.93% in TREC 4, 0.99% in TREC 6 and 0.42%

in TREC 1. This small boost in performance proves that the proposed GVSM model is promising. There are many aspects though in the GVSM that we think require further investigation, like for example the fact that we have not conducted WSD so as to map each document and query term occurrence into its correct sense, or the fact that the weighting scheme of the edges used in SR generates from the distribution of each edge type in WordNet, while there might be other more sophisticated ways to compute edge weights. We believe that if these, but also more aspects discussed in the next section, are tackled, the proposed GVSM may improve more the retrieval performance.

5 Future Work

From the experimental evaluation we infer that SR performs very well, and in fact better than all the tested related measures. With regards to the GVSM model, experimental evaluation in three TREC collections has shown that the model is promising and may boost retrieval performance more if several details are further investigated and further enhancements are made. Primarily, the computation of the maximum semantic relatedness between two terms includes the selection of the semantic path between two senses that maximizes SR. This can be partially unrealistic since we are not sure whether these senses are the correct senses of the terms. To tackle this issue, WSD techniques may be used. In addition, the role of phrase detection is yet to be explored and
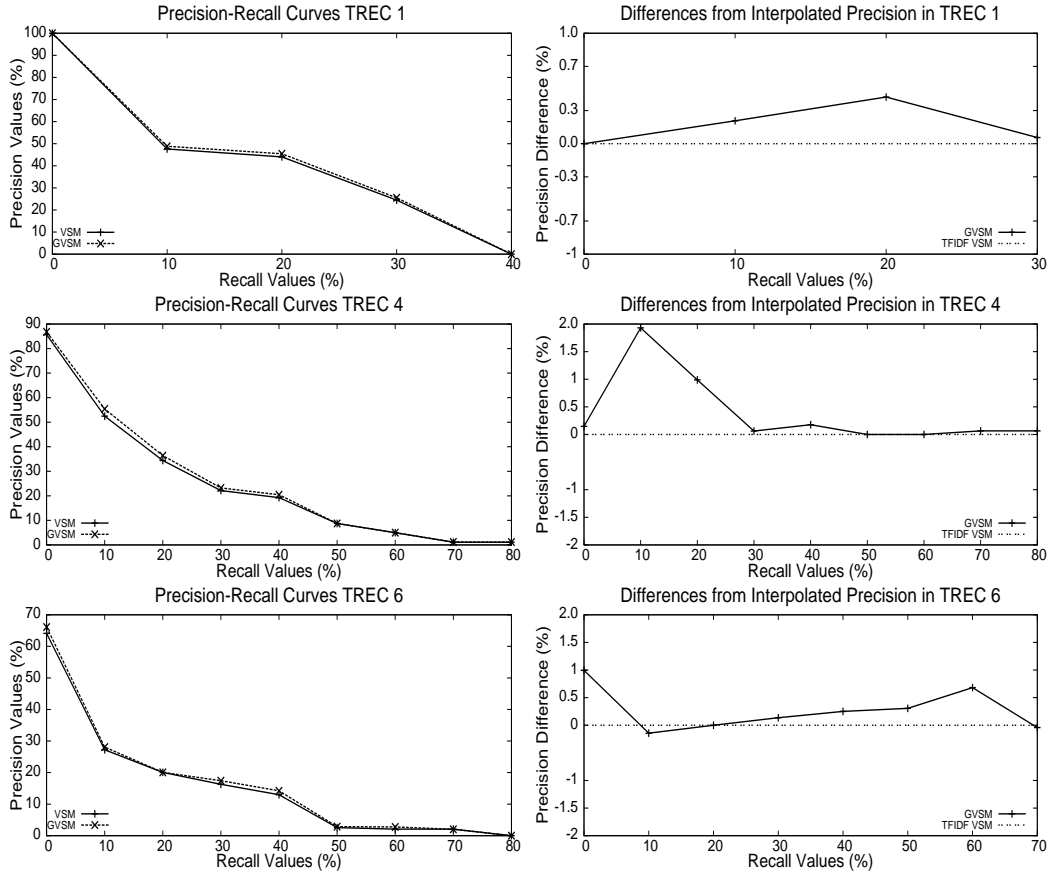
Figure 3: Differences (%) from the baseline in interpolated precision.

added into the model. Since we are using a large knowledge-base (WordNet), we can add a simple method to look-up term occurrences in a specified window and check whether they form a phrase. This would also decrease the ambiguity of the respective text fragment, since in WordNet a phrase is usually monosemous.

Moreover, there are additional aspects that deserve further research. In previously proposed GVSM, like the one proposed by Voorhees (1993), or by Mavroeidis et al. (2005), it is suggested that semantic information can create an individual space, leading to a dual representation of each document, namely, a vector with document's terms and another with semantic information. Rationally, the proposed GVSM could act complementary to the standard VSM representation. Thus, the similarity between a query and a document may be computed by weighting the similarity in the terms space and the senses' space. Finally, we should also examine the perspective of applying the proposed measure of semantic relatedness in a query expansion technique, similarly to the work of Fang (2008).

6 Conclusions

In this paper we presented a new measure of semantic relatedness and expanded the standard VSM to embed the semantic relatedness between pairs of terms into a new GVSM model. The semantic relatedness measure takes into account all of the semantic links offered by WordNet. It considers WordNet as a graph, weighs edges depending on their type and depth and computes the maximum relatedness between any two nodes, connected via one or more paths. The comparison to well known measures gives promising results. The application of our measure in the suggested GVSM demonstrates slightly improved performance in information retrieval tasks. It is on our next plans to study the influence of WSD performance on the proposed model. Furthermore, a comparative analysis between the proposed GVSM and other semantic network based models will also shed light towards the conditions, under which, embedding semantic information improves text retrieval.

References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley.

S. Brody, R. Navigli, and M. Lapata. 2006. Ensemble methods for unsupervised wsd. In *Proc. of COLING/ACL 2006*, pages 97–104.

A. Budanitsky and G. Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

T.H. Cormen, C.E. Leiserson, and R.L. Rivest. 1990. *Introduction to Algorithms*. The MIT Press.

H. Fang. 2008. A re-examination of query expansion using lexical resources. In *Proc. of ACL 2008*, pages 139–147.

C. Fellbaum. 1998. *WordNet – an electronic lexical database*. MIT Press.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2002. Placing search in context: The concept revisited. *ACM TOIS*, 20(1):116–131.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of the 20th IJCAI*, pages 1606–1611. Hyderabad, India.

G. Hirst and D. St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In *WordNet: An Electronic Lexical Database, chapter 13*, pages 305–332, Cambridge. The MIT Press.

M. Jarmasz and S. Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proc. of Conference on Recent Advances in Natural Language Processing*, pages 212–219.

J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of ROCLING X*, pages 19–33.

R. Krovetz and W.B. Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

C. Leacock, G. Miller, and M. Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, March.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*, pages 296–304.

D. Mavroeidis, G. Tsatsaronis, M. Vazirgiannis, M. Theobald, and G. Weikum. 2005. Word sense disambiguation for exploiting hierarchical thesauri in text classification. In *Proc. of the 9th PKDD*, pages 181–192.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant word senses in untagged text. In *Proc, of the 42nd ACL*, pages 280–287. Spain.

R. Mihalcea and A. Csomai. 2005. Senselearner: Word sense disambiguation for all words in unrestricted text. In *Proc. of the 43rd ACL*, pages 53–56.

R. Mihalcea, P. Tarau, and E. Figa. 2004. Pagerank on semantic networks with application to word sense disambiguation. In *Proc. of the 20th COLING*.

G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

A. Montoyo, A. Suarez, G. Rigau, and M. Palomar. 2005. Combining knowledge- and corpus-based word-sense-disambiguation methods. *Journal of Artificial Intelligence Research*, 23:299–330, March.

P. Resnik. 1995. Using information content to evaluate semantic similarity. In *Proc. of the 14th IJCAI*, pages 448–453, Canada.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

M. Sanderson. 1994. Word sense disambiguation and information retrieval. In *Proc. of the 17th SIGIR*, pages 142–151, Ireland. ACM.

R. Sinha and R. Mihalcea. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proc. of the IEEE International Conference on Semantic Computing*.

C. Stokoe, M.P. Oakes, and J. Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proc. of the 26th SIGIR*, pages 159–166.

M. Strube and S.P. Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *Proc. of the 21st AAAI*.

G. Tsatsaronis, M. Vazirgiannis, and I. Androutsopoulos. 2007. Word sense disambiguation with spreading activation networks generated from thesauri. In *Proc. of the 20th IJCAI*, pages 1725–1730.

E. Voorhees. 1993. Using wordnet to disambiguate word sense for text retrieval. In *Proc. of the 16th SIGIR*, pages 171–180. ACM.

S.K.M. Wong, W. Ziarko, V.V. Raghavan, and P.C.N. Wong. 1987. On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems*, 12(2):299–321.