Unsupervised Semantic Similarity Computation Between Terms Using Web Documents

Elias Iosif, *Student Member, IEEE* and Alexandros Potamianos, *Member, IEEE*

***Abstract*— In this work, web-based metrics for semantic similarity computation between words or terms are presented and compared with the state-of-the-art. Starting from the fundamental assumption that similarity of context implies similarity of meaning, context-based metrics use a web search engine to download relevant documents and then exploit the retrieved contextual information for the words of interest. The proposed algorithms can be generalized and applied to other languages, work automatically and do not require any human annotated knowledge resources, e.g., ontologies. Context-based metrics are evaluated on the Charles-Miller dataset and on a medical term dataset. It is shown that the context-based similarity metrics significantly outperform co-occurrence based metrics, in terms of correlation with human judgment, for both tasks. In addition, the proposed context-based algorithms are shown to be competitive with state-of-the-art supervised semantic similarity metrics that employ language-specific knowledge resources. Specifically, context-based metrics achieve correlation scores of up to 0.88% and 0.74% for the Charles-Miller and medical datasets, respectively. In the framework of contextual metrics we discuss several weighting schemes for feature selection with respect to the different nature of the experimental corpora. Also, an additional step towards better understanding of feature saliency for similarity estimation, is presented, considering different types of contextual features. In final, we outline a first attempt for the computation of document "grammaticality", which tries to take into account the diversity in language use of the downloaded documents.**

***Index Terms*— Natural language processing, semantic similarity**

I. Introduction

NUMEROUS information retrieval and natural language processing applications require knowledge of semantic similarity between words or terms. For example, in query expansion the addition of semantically similar words to the original query is likely to increase the relevance of retrieved documents [1]. Specifically, in [2], [3], [4], it is shown that query expansion using related words acquired from WordNet increases the recall of retrieved documents. Moreover, semantic similarity measures are important for many natural language processing (NLP) tasks, such as language modeling [5], grammar induction [6], word sense disambiguation [7], speech understanding and spoken dialogue systems [5]. In [8], [9], several unsupervised, statistical metrics are discussed for automatic induction of semantic classes, applied on homogeneous and heterogeneous corpora.

Elias Iosif is with the Dept. of Electronics & Computer Engineering, Technical Univ. of Crete, Chania 73100, Greece; email: iosife@telecom.tuc.gr; tel:+30-28210-37368; fax:+30-28210-37542.

Alexandros Potamianos is with the Dept. of Electronics & Computer Engineering, Technical Univ. of Crete, Chania 73100, Greece; email: potam@telecom.tuc.gr; tel:+30-28210-37221; fax:+30-28210-37542.

Early versions of this work were partially supported by the EU-IST-FP6 MUSCLE network of excellence.

The majority of the semantic similarity metrics employed today use hand-crafted language resources [10], [11], [12], [13]. The use and updating of resources, such as thesauri or ontologies, is a time consuming and tedious task, demanding human labor and often expert knowledge. Also, language resources are not ubiquitous and are unavailable for many languages. As a result these methods are of little utility for applications where human and language resources are sparse. In addition, these methods cannot be applied for words or terms that are not included in the resource repository, e.g., scientific terms, out of vocabulary words, neologisms. To overcome this problem knowledge resources are often constructed for specific domains where general purpose ontologies do not offer adequate term coverage. For example, in addition to WordNet, domain-specific ontologies, e.g., MeSH, is used for applications in the (bio)medical domain [10]. Improving term coverage remains an open research issue; algorithms are proposed in the literature on how to pool multiple knowledge resources or add terms to existing language resources, e.g., ontology merging techniques and cross-ontology similarity metrics. The work of Budanitsky and Hirst [14] provides a thorough review of different metrics that use WordNet for computing semantic similarity.

The web has a multilingual character; new words, neologisms and occasionalisms (hapax legomena), are added frequently and efficiently. Thus, it is the obvious place for mining semantic relationships for unseen words. Also, the web contains both general purpose words, found in news articles and blogs, as well as, scientific terminology, found in documents written by experts. Overall, the web covers a plethora of domains, authoring styles and languages, and is fertile ground for automatic semantic knowledge acquisition. The web is a valuable resource that has been exploited for a variety of NLP applications. In [15], the web page counts returned by a search engine are used to estimate the probability of n-gram language models. In [16], the web page counts of fixed lexical patterns are used to identify synonymy and antonymy between nouns. An extension of this approach was proposed in [17]; web queries of lexico-syntactic patterns were used for discovering relationships between verbs. The web is also an invaluable source for constructing text corpora. For example, in [18], a large corpus of web pages was constructed and used for word sense disambiguation. Other applications were automatically-constructed web corpora have been used to train statistical models include statistical machine translation [19] and question-answering systems [20].

Recently there has been much research interest in developing web-based similarity measures. Typically such approaches use the results returned by one or more web search engines

using one or multiple queries. Web-based similarity measures can broadly divided into three categories: (i) measures that rely only on the number of the returned hits, (ii) measures that download a number of the top-ranked documents and then apply text processing techniques, (iii) measures that combine the two approaches. Web-based similarity computation algorithms are used in a diverse range of applications, such as automatic annotation of web pages [21], social networks construction [22], [23] and music genre classification [24], [25]. However, in most cases, the form of the web query and/or the feature extraction process is application-dependent, e.g., if one is interested in movie genre classification it is useful to include the term "movie" in the submitted query.

In this work, we focus on the problem of fully unsupervised web-based semantic similarity computation between words or terms; no hand-crafted rules or resources are employed. Web search engines are used for text corpus mining and context-based similarity distances are automatically computed on this corpus. The proposed algorithm requires no expert knowledge or language resources, and is thus mitigates the need for language-dependent adaptation. In order to calculate the semantic similarity between words, we investigate two families of unsupervised, web-based similarity metrics [1]. The first type considers only the number of hits returned by a web search engine, as in [26] and [27]. The second is fully text-based, downloads the top-ranked documents returned by a web query and compares the context around words of interest to estimate semantic similarity. This work extends our work in [28]. The following are the original contributions of this work compared to [28]:

1) Several new contextual similarity algorithms are proposed and evaluated over large collections of downloaded documents.
2) The metrics are evaluated both on the Charles-Miller and on a medical term dataset, i.e., in this work we investigate both word and term similarity. The two evaluation domains are also semantically different: ordinary words of general use vs medical terms.
3) We demonstrate the effect of feature and document selection on semantic similarity computation. For example, it is shown that non-content words (stop-words) are important features for word similarity computation but poor features for term similarity computation.
4) We show that the proposed fully unsupervised method based on context similarity can compete with state-of-the-art supervised similarity metrics that employ elaborate language resources.

The evaluation results also provide insight into human cognition and especially the language acquisition process, which is also (at the semantic level) mostly unsupervised.

The remainder of this work is organized as follows. In Section II, an overview of related work in the area of semantic similarity computation is presented. Page-count based similarity metrics are defined in Section III. The proposed context-based similarity algorithms are presented in Section IV. In Section V, the semantic similarity computation algorithm is described along with the experimental procedure. In Section VI, the evaluation results are reported for the proposed algorithms for two evaluation datasets. The results are compared with state-of-the-art semantic similarity algorithms that employ knowledge resources such as WordNet and MeSH. The results are further discussed in Section VII, and implications of feature selection and document selection for context-based similarity metrics are presented. Finally, the work concludes with Section VIII where promising directions for further research are also given.

II. Related work

Metrics that measure semantic similarity between words or terms can be classified into four main categories depending if knowledge resources are used or not: (a) supervised *resource-based metrics*, consulting only human-built knowledge resources, such as ontologies, (b) supervised *knowledge-rich text-mining metrics*, i.e., metrics that perform text mining relying also on knowledge resources, (c) unsupervised *co-occurrence metrics*, i.e., unsupervised metrics that assume that the semantic similarity among words or terms can be expressed by an association ratio which is a function their co-occurrence and (d) unsupervised *text-based metrics*, i.e., metrics that are fully text-based and exploit the context or proximity of words or terms to compute semantic similarity. The last two categories of metrics do not use any language resources or expert knowledge, both depend only on web search engines. In this sense, the metrics are referred to as "unsupervised"; no semantically labeled human-annotated data is required to compute the semantic distance between words or terms. Resource-based and knowledge-rich text-mining metrics, however, use such data, and are henceforth referred to as "supervised" metrics.

Several resource-based methods have been proposed in the literature that use, e.g., WordNet, for semantic similarity computation. Edge counting methods consider the length of the paths that links the words, as well as the word positions in the taxonomic structure [11], [12]. Information content methods compute similarity between words by combining taxonomic features that exist in the used resource, e.g., number of subsumed words, with frequencies computed over textual corpora [13]. Hybrid methods combine synsets [2] with word neighborhoods and other features [10]. In the work of Bollegala et al. [26], a hybrid method, among others, is defined that combines page counts, returned by a search engine, and lexico-syntactic patterns, extracted from the returned snippets using a number of synonymous nouns acquired from WordNet.

Co-occurrence-based metrics attempt to implement computational models for the notion of "word association" which is used in psycholinguistics. This notion describes the procedure of lexical decision of human associative memory. In [29], an association ratio is proposed using the information theoretic

[1]It should be noted that the used similarity measures are not true metrics (distance functions), since they do not satisfy the triangle inequality, i.e., for words or terms x, y, z: $S(w_x, w_z) \not\leq S(w_x, w_y) + S(w_y, w_z)$, where $S(.)$ denotes the similarity measure. However, in this work the we use the notions of "metric" and "measure" interchangeably.

[2]A synset is a set of words (or terms) that are considered to be synonymous. This notion is widely used in lexical resources like WordNet.

metric of mutual information in order to identify patterns which can be used for the construction of semantic classes. In [26], several association metrics are applied, using a search engine in order to obtain co-occurrence counts for a word pair. If the pair of interest consists of the words w_1 and w_2, their co-occurrence frequency is taken to be equal to the number of hits returned by a search engine, given a query of the form "w_1 AND w_2".

Text-based metrics typically use contextual features to compute semantic similarity. Context-based metrics operate under the assumption that words with similar contexts have similar meaning. One of the first studies of this hypothesis is the work of Rubenstein and Goodenough stating that "words that are similar in meaning occur in similar contexts" [30]. Using this assumption the semantic similarity between two words can be estimated by measuring the difference between the probability distributions of their contextual features. Various context-based metrics have been proposed in the literature, such as: Kullback-Leibler, Information-radius and Manhattan norm [6], [8]. The contextual probability distributions can be estimated (and smoothed) according to n-gram language models [31]. Another representation of the contextual environment of a word is the *bag-of-words* model [32]. According to this model, the contextual features of a word form the elements of a vector. Assuming independence among the features, the similarity of two words is computed as the product of their feature vectors (cosine similarity) [33], [9]. More recently, context-based similarity metrics construct document collections by querying web search engines and downloading a number of the returned top-ranked documents, in order to compute semantic similarity between words or terms [28].

III. Page-count-based similarity metrics

Page-count-based metrics use association ratios between words that are computed using their co-occurrence frequency in documents. The basic assumption of this approach is that high association ratios indicate a semantic relations between words[3] [29]. For the documents indexed by a search engine we define the notations shown in Table I [34]. Four co-occurrence measures are evaluated in this work, namely, Jaccard coefficient, Dice coefficient and Mutual Information, as in [26], and Google-based Semantic Relatedness [27], to compute semantic similarity between word pairs.

Notation	Description
$\{D\}$	set of all documents indexed by search engine
$\lvert D \rvert$	number of documents in $\{D\}$
w	a word or term
$\{D \mid w\}$	subset of $\{D\}$, documents indexed by w
$\{D \mid w_1, w_2\}$	subset of $\{D\}$, documents indexed by w_1 and w_2
$\lvert D \mid w \rvert$	fraction of documents in $\{D\}$ indexed by w
$\lvert D \mid w_1, w_2 \rvert$	fraction of documents in $\{D\}$ indexed by w_1 and w_2

TABLE I

Definitions for document sets indexed by search engines.

A. Jaccard and Dice coefficient

The Jaccard coefficient is a measure calculating the similarity (or diversity) between sets. We use a variation of the Jaccard coefficient in this work, defined as:

$$J(w_1, w_2) = \frac{\lvert D \mid w_1, w_2 \rvert}{\lvert D \mid w_1 \rvert + \lvert D \mid w_2 \rvert - \lvert D \mid w_1, w_2 \rvert} \quad (1)$$

In probabilistic terms, Eq. 1 finds the maximum likelihood estimate of the ratio of the probability of finding a document where words w_1 and w_2 co-occur over the probability of finding a document where either w_1 or w_2 occurs[4]. If w_1 and w_2 are the same word then the Jaccard coefficient is equal to 1 (absolute semantic similarity). If two words never co-occur in a document then the Jaccard coefficient is 0. The Dice coefficient is related to the Jaccard coefficient and is computed as:

$$C(w_1, w_2) = \frac{2 \lvert D \mid w_1, w_2 \rvert}{\lvert D \mid w_1 \rvert + \lvert D \mid w_2 \rvert} \quad (2)$$

Again, the Dice coefficient is equal to 1 if w_1 and w_2 are identical, and 0 if two words never co-occur.

B. Mutual information

If we consider that the occurrence of words w_1 and w_2 is a random variables X and Y, respectively, then the pointwise mutual information (MI) among X and Y measures the mutual dependence between the appearance of words w_1 and w_2 [29]. The maximum likelihood estimate of MI is

$$I(X, Y) = \log \frac{\frac{\lvert D \mid w_1, w_2 \rvert}{\lvert D \rvert}}{\frac{\lvert D \mid w_1 \rvert}{\lvert D \rvert} \frac{\lvert D \mid w_2 \rvert}{\lvert D \rvert}} \quad (3)$$

Mutual information measures the information that variables X and Y share. It quantifies how the knowledge of one variable reduces the uncertainty about the other. For instance, if X and Y are independent, then knowing X does not give any information about Y and the mutual information is 0. For $X = Y$, the knowledge of X gives the value of Y with certainty and the mutual information is 1. Note that the fractions of documents are normalized by the number of documents indexed by the search engine, $\lvert D \rvert$, giving a maximum likelihood estimate of the probability of finding a document in the web that contains this word.

C. Google-based Semantic Relatedness

Motivated by Kolmogorov complexity, Cilibrasi and Vitanyi [35], [36] proposed a page-count-based similarity measure, called the Normalized Google Distance [5], defined as

$$G(w_1, w_2) = \frac{\max\{A\} - \log \lvert D \mid w_1, w_2 \rvert}{\log \lvert D \rvert - \min\{A\}}, \quad (4)$$

[3]It is interesting to note that web-based co-occurrence metrics often outperform more elaborate corpus-based metrics. This shows that overcoming the data sparseness problem is sometimes more important than building an accurate estimator. For example an improved n-gram language probability estimation using web n-gram occurrence can be found in the literature [15].

[4]Normalization by the total number of documents $\lvert D \rvert$ is the same for the nominator and denominator, and can be ignored.

[5]Note that all metrics presented in this work use search engines, thus, the term "Google" to refer to this specific metric can be somewhat misleading.

where $A = \{\log \mid D \mid w_1 \mid, \log \mid D \mid w_2 \mid\}$. As the semantic similarity between two words increases the distance computed by Eq. 4 decreases. Thus, this metric can be considered as a dissimilarity measure. Note that it is unbounded, ranging from 0 to ∞. In [27], a variation of Normalized Google Distance is proposed in order to define a similarity measurement. This variation is called Google-based Semantic Relatedness:

$$G^{'}(w_1, w_2) = e^{-2G(w_1, w_2)} \tag{5}$$

where $G(w_1, w_2)$ is computed according to Eq. 4. Note that the Google-based Semantic Relatedness is bounded taking values between 0 and 1.

IV. Text-based similarity metrics

In this section, a family of text-based similarity metrics is introduced that computes cosine similarity between feature vectors extracted from word or term context, i.e., a "bag-of-words" context model. The basic assumption behind this metrics is that *similarity of context implies similarity of meaning*, i.e., words that appear in similar lexical environment (left and right contexts), have a close semantic relation [30], [6], [8].

"Bag-of-words" [37], [9] models assume that the feature vector consists of words whose occurrence in text is independent of each other. The proposed context-based metrics employ a context window of fixed size K words for feature extraction. Specifically, for each occurrence of a word of interest w in the corpus, the right and left contexts of length K are considered, e.g., $[v_{K,L}\ ...\ v_{2,L}\ v_{1,L}]\ w\ [v_{1,R}\ v_{2,R}\ ...\ v_{K,R}]$, where $v_{i,L}$ and $v_{i,R}$ represent the i^{th} word to the left and to the right of w respectively. The feature vector for every word w is defined as $T_{w,K} = (t_{w,1}, t_{w,2}, ..., t_{w,N})$, where $t_{w,i}$ is a non-negative integer and K is the context window size. Note that the feature vector size is equal to the vocabulary size N, i.e., we have a feature for each word in the vocabulary V. The i^{th} feature value $t_{w,i}$ reflects the (frequency of) occurrence of vocabulary word v_i within the left or right context window K of word (or term) w. The feature value $t_{w,i}$ may be set according to a variety of schemes that take into account the frequency of occurrence of a feature. Once the feature weighting scheme is selected the "bag-of-words"-based metric S^K computes the similarity between two words or terms, w_1 and w_2, as the cosine similarity of their corresponding feature vectors, $T_{w_1,K}$ and $T_{w_2,K}$ [33], [9]:

$$S^K(w_1, w_2) = \frac{\sum_{i=1}^{N} t_{w_1,i} t_{w_2,i}}{\sqrt{\sum_{i=1}^{N} (t_{w_1,i})^2} \sqrt{\sum_{i=1}^{N} (t_{w_2,i})^2}} \tag{6}$$

given a context window of length K. The cosine similarity metric assigns 0 similarity score to completely dissimilar words, and 1 for identical words.

The various feature weighting schemes for computing $t_{w,i}$ used in this work are presented in Table II. The weighting schemes can be classified into binary and frequency-based. The binary metric assigns weight $t_{w,i} = 1$ when the i^{th} word in the vocabulary exists at the left or right context of at least one instance of the word w, and 0 otherwise. Frequency-based weighting schemes compute the (normalized) frequency of occurrence of context words. Various frequency-based weighting schemes popular in natural language processing and web applications are proposed and evaluated, specifically, term-frequency (TF), logarithmic term-frequency (LTF), term frequency inverse document frequency (TFIDF), logarithmic TFIDF, and add-one smoothing of these methods. The value of the weight $t_{w,i}$ for the frequency-based metrics are shown in Table II, where $c(v_i)$ denotes the number of occurrences of the i^{th} word in the vocabulary v_i within the left or right context of all occurrences of word w in a corpus. Note that the context frequency of v_i is normalized by the number of occurrences of word w in the corpus, i.e., $c(w)$. For the case of add-one-metrics, α denotes the cardinality of set, which contains the unique words that appear in the context(s) of w.

Scheme	Value of $t_{w,i}$ for $c(v_i) > 0$
Binary (B)	1
Term frequency (TF)	$\frac{c(v_i)}{c(w)}$
Log of TF (LTF)	$\frac{\log(c(v_i))}{\log(c(w))}$
Add-one LTF ($LTF1$)	$\frac{\log(c(v_i)+1)}{\log(c(w)+\alpha)}$
TF-inverse document frequency ($TFIDF$)	$\frac{c(v_i)}{c(w)} \log \frac{\lvert D \rvert}{\lvert D \rvert v_i}$
Log of $TFIDF$ ($LTFIDF$)	$\frac{\log(c(v_i))}{\log(c(w))} \log \frac{\lvert D \rvert}{\lvert D \rvert v_i}$
Add-one $LTFIDF$ ($LTF1IDF$)	$\frac{\log(c(v_i)+1)}{\log(c(w)+\alpha)} \log \frac{\lvert D \rvert}{\lvert D \rvert v_i}$

TABLE II
Context Feature Weighting Schemes

Logarithmic term frequency (LTF) weighting is similar to term frequency (TF), the main difference being the non-linear scaling of counts and the assignment of weight 0 to singletons, i.e., context words appearing only once. Applying logarithmic scheme, the highly frequent contextual features are not allowed to dominate the computation of similarity score, as happens in the case of term frequency scheme. Also, the non-linearity introduced by the logarithmic scheme it is a simple way to approach the non-linear process by which the human memory builds the semantic associations between words, assuming that the contextual features are taken into account during the cognitive process. Logarithmic add-one smoothing (LTF1) takes singletons into account with a positive weight of $\frac{\log(2)}{\log(c(w)+\alpha)}$.

The term-frequency inverse document frequency (TFIDF) metric is a popular metric in information retrieval that assigns more weight to semantically salient words, effectively reducing the effect of stop words and non-content words. Similarly, in this work, the logarithm of the inverse document frequency of context words in each document is computed as $\log \frac{|D|}{|D|v_i|}$ and used to multiply the TF estimate. Similarly the logarithmic TFIDF (LTFIDF) multiplies the LTF estimate with the inverse document frequency. Finally, the add-one smoothing version of this metric is computed (LTF1IDF).

V. Corpus based similarity computation

We experimented with (i) page-count-based, and (ii) text-based similarity metrics, described in Section III and Section IV, respectively. For page-count metrics the Yahoo! search engine was used to determine the frequency occurrence and

co-occurrence of two words w_1 and w_2. Specifically, the total number of hits for the queries "w_1", "w_2" and "w_1 AND w_2" were used to compute the Jaccard, Dice, Mutual Information and Google metrics.

For the contextual similarity metrics, for each pair of words or terms, "w_1 w_2" a few hundreds of documents were downloaded using "w_1 AND w_2" (e.g., "boy AND lad") queries. The *URLs* for the top ranked documents were retrieved using the Yahoo! search engine via the Yahoo! Search API, which is freely available [38]. This query type retrieves documents containing both words, as opposed to generic "w_1 OR w_2" queries that download documents containing either word. In [28], preliminary experiments have shown that AND queries significantly outperform OR queries for context-based semantic similarity computation. Once the documents are downloaded, the left and right contexts of all occurrences of w_1 and w_2 are examined and the corresponding feature vectors are constructed according to the following experimental parameters:

1) Number of web documents, $\mid D \mid$: how many web documents are used.
2) Contextual window size, K: the left and right contexts of w_1 and w_2 are examined according to the value of contextual window size. The window size is applied within the sentence boundaries.
3) Stop words filtering (yes/no): consideration (or not) of stop words in the feature vectors.
4) Type of weighting scheme: the values of vector features are set according to one of the weighting schemes presented in Table II.

The semantic similarity between w_1 and w_2 is then computed as the cosine similarity between the feature vectors as shown in Eq. 6.

VI. Evaluation

In this section, we present a comparative evaluation of the proposed similarity algorithms, in terms of correlation, with respect to the human ratings of: (i) Charles-Miller dataset of common words, and (ii) the MeSH dataset of medical terms. Both the page-count-based similarity metrics defined in Section III are evaluated, as well as, the proposed fully text-based similarity algorithms defined in Section IV. The proposed algorithms are also compared with metrics that use knowledge resources, e.g., the WordNet ontology for the Charles-Miller dataset, and the MeSH ontology for the MeSH dataset of medical terms.

A. Corpus description

For evaluation purposes we used two datasets: (a) the commonly used Charles-Miller dataset of common words [39], and (b) a dataset of medical terms included in the MeSH ontology. The first dataset consists of 28 noun pairs of general use that were rated according to their semantic similarity by 38 human subjects. The assigned similarity scores range from 0 (not similar) to 4 (perfect synonymy). The selection of this dataset was mainly motivated by its wide use that enabled us to compare our work with a variety of other approaches.

The MeSH dataset includes 34 medical terms pairs that have been rated for similarity by experts. MeSH is the acronym for "Medical Subject Headings" and is a taxonomic hierarchy containing medical terms proposed by the National Library of Medicine of USA. The full MeSH dataset [6] contains 36 pairs of MeSH terms, such as, "asthma-pneumonia" and "anemia-appendicitis". In this work, we used a subset of 34 pairs due to the limited amount of web documents available for the 5^{th} and 36^{th} pair. The MeSH dataset along with the human-rated similarity scores were taken from the work of Petrakis et al. [10]. Petrakis et al. asked Dr. Qi at Dalhousie University to construct a set of MeSH term pairs. Then medical experts were asked to submit similarity scores for the MeSH term pairs using a web-based tool. In total, 8 experts took part in the above procedure assigning similarity scores from 0 (no similarity) to 4 (absolute similarity). Pairs with standard deviation of similarity scores higher than the user defined threshold of $t = 0.8$ were excluded from the evaluation. The MeSH dataset was selected in order to investigate similarity between terms that are rated by experts rather than naive subjects.

Let $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ be the random vectors that store the similarity scores given by human subjects and the computational metric, respectively, for each of the $i = 1, 2, ..., n$ word pairs. The correlation coefficient between the scores produced by humans and machine is estimated using the Pearson correlation, as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, for $i = 1, 2, ..., n$.

B. Evaluation of page-count-based metrics

The correlation scores between the page-count-based semantic similarity metrics and human ratings are presented in Table III for the two tasks. The similarity metrics based on

Dataset	J	C	I	G'
Charles-Miller	0.41	0.41	0.69	0.66
MeSH	0.26	0.29	0.30	0.41

TABLE III
Correlation of page-count metrics.

the Jaccard (J) and Dice (C) coefficients achieve comparable correlation performance, which is expected given the similarities between the two metrics. The Mutual Information (I) and Google-based Semantic Relatedness (G') achieve significantly [7] better performance than Jaccard and Dice, especially for the case of Charles-Miller dataset. Overall, the achieved correlation for the words of general use in the Charles-Miller dataset is significantly greater compared to the medical terms of MeSH dataset, for all metrics.

[6] http://www.intelligence.tuc.gr/similarity/datasets/MeSHDataset.pdf.

[7] Wherever we compare metrics that were implemented in this work, stating that their difference is significantly better, the term "significantly" refers to statistical significance with significance level greater than 95%. For significance test we used paired t-test.

C. Evaluation of context-based metrics

Next we present the performance of the context-based metrics for the various feature weighting schemes shown in Table II and for different contextual window sizes K. The performance of each metric is shown as a function of the number of downloaded documents. The correlation scores for the Charles-Miller dataset are shown in Fig. 1(a) and (b), and for the MeSH dataset in Fig. 1(c) and (d).

In Fig. 1(a), the correlation scores for the Charles-Miller dataset are shown using several weighting schemes. Performance is shown as a function of the context window K (ranging from 1 to 20) for a total number of 100 downloaded documents. For most metrics, highest correlation is achieved with context size $K = 1$, i.e., considering only the immediate context of one word to the left and one to the right. For larger context windows, performance degrades fast especially for the TFIDF weighting schemes. The highest correlation score of correlation 0.72 is achieved by the LTF scheme with the binary weighting scheme being a close second. Note that the linear frequency-based weighting schemes, i.e., TF and TFIDF, perform poorly, compared to their logarithmic counterparts, especially, for large context sizes.

In Fig. 1(b), the performance of the binary (B), LTF and LTFIDF weighting schemes are shown for a context window size of $K=1$ as a function of the number of downloaded documents (ranging from 10 to 1000). The correlation improves with the number of documents and the performance bound is not reached even at 1000 documents. Good correlation performance is reached even with as few as 30 documents, however, it is clear that the performance of the similarity metric is not robust if fewer than 100 documents are used. Overall, the LTF scheme performs the best up to approximately 500 documents while the binary scheme provides best performance for a larger number of documents. Also note, that the performance gap between LTF and LTFIDF is bridged for a large number of documents. Overall, the highest correlation score of 0.88 is achieved using the binary weighting scheme and 1000 documents.

In Fig. 1(c), the correlation score for the MeSH dataset is shown. The weighting schemes, context window size and number of documents (100) are the same as in (a), and thus the two plots are directly comparable. The main differences in performance for the MeSH dataset compared to the Charles-Miller dataset are: (i) the relative performance of the weighting schemes, i.e., for the MeSH dataset the LTFIDF weighting scheme significantly outperforms all other schemes (note the especially poor performance of the binary weighting scheme), and (ii) the optimum context window size, i.e., for the MeSH dataset best correlation scores are achieved for context window size between $K=2$ and $K=5$, as opposed to $K=1$ for the Charles-Miller. In addition, the degradation of performance for large context windows is much more graceful for the MeSH dataset. The best correlation score is 0.67 for $K=3$ and the LTFIDF weighting scheme.

In Fig. 1(d), the performance of the binary (B), LTF, and LTFIDF weighting schemes are shown as a function of the number of downloaded documents. The window size used is $K=1$ in order for plots (b) and (d) to be directly comparable[8]. As in (b), correlation increases as more documents are considered. However, in (d), the highest score is achieved for 800 documents and the performance degrades somewhat for 1000 documents. The LTFIDF weighting scheme significantly outperforms the other two metrics, while the binary scheme performs the worst, i.e., the relative metric performance is reversed in (d) compared to (b). Finally, note that the absolute performance of the metrics for the MeSH dataset is worse than for the Charles-Miller dataset; this is consistent with results reported in the literature.

Add-one smoothing schemes LTF1 and LTF1IDF that do not discard contextual singletons achieve almost identical correlation scores to LTF and LTFIDF respectively. Thus, the results for LTF1 and LTF1IDF are not included in the plots.

D. Stop-word filtering

Motivated by the differences in performance between the term weighting schemes for the word and term tasks, we investigate next how stop-word filtering affects performance. For this purpose we classified the contextual words into stop-words (sw) and non-stop-words and computed the semantic similarity scores for the three possible setups, i.e., (i) only stop-words are considered in the similarity computation algorithm, (ii) any word not being a stop-word is considered, and (iii) any word is considered (same setup as that used for Fig. 1 (a),(c)). The correlation score was computed for 100 documents using context window size $K = 1$. For each dataset the best weighting scheme was used, i.e., LTF for Charles-Miller and LTFIDF for MeSH. The results are shown in Table IV.

Dataset	Type of context		
	only stop-words (sw)	w/o sw	both
Charles-Miller	0.68	0.64	0.72
MeSH	0.25	0.66	0.63

TABLE IV
CORRELATION FOR DIFFERENT TYPES OF CONTEXT.

For the Charles-Miller dataset, the inclusion of stop-words boosts overall performance, in fact, similarity computation using only stop-words as features outperforms somewhat similarity computation using only non stop-words! For the MeSH terms dataset, however, stop-word-based similarity computation performs very poorly. In fact, including stop-words seems to be hurting overall performance; from 0.66 when stop-words are excluded to 0.63. For a more detailed discussion on stop-word filtering and feature selection see Section VII.

E. Unsupervised vs supervised metrics

Next the performance of the proposed unsupervised algorithms are compared with semantic similarity computation algorithms found in the literature. In addition to page-count similarity metrics, we also consider metrics that consult

[8]Although for 100 documents the best correlation score is obtained for $K=3$, for large number of documents comparable performance is obtained for context window sizes of one, two or three.

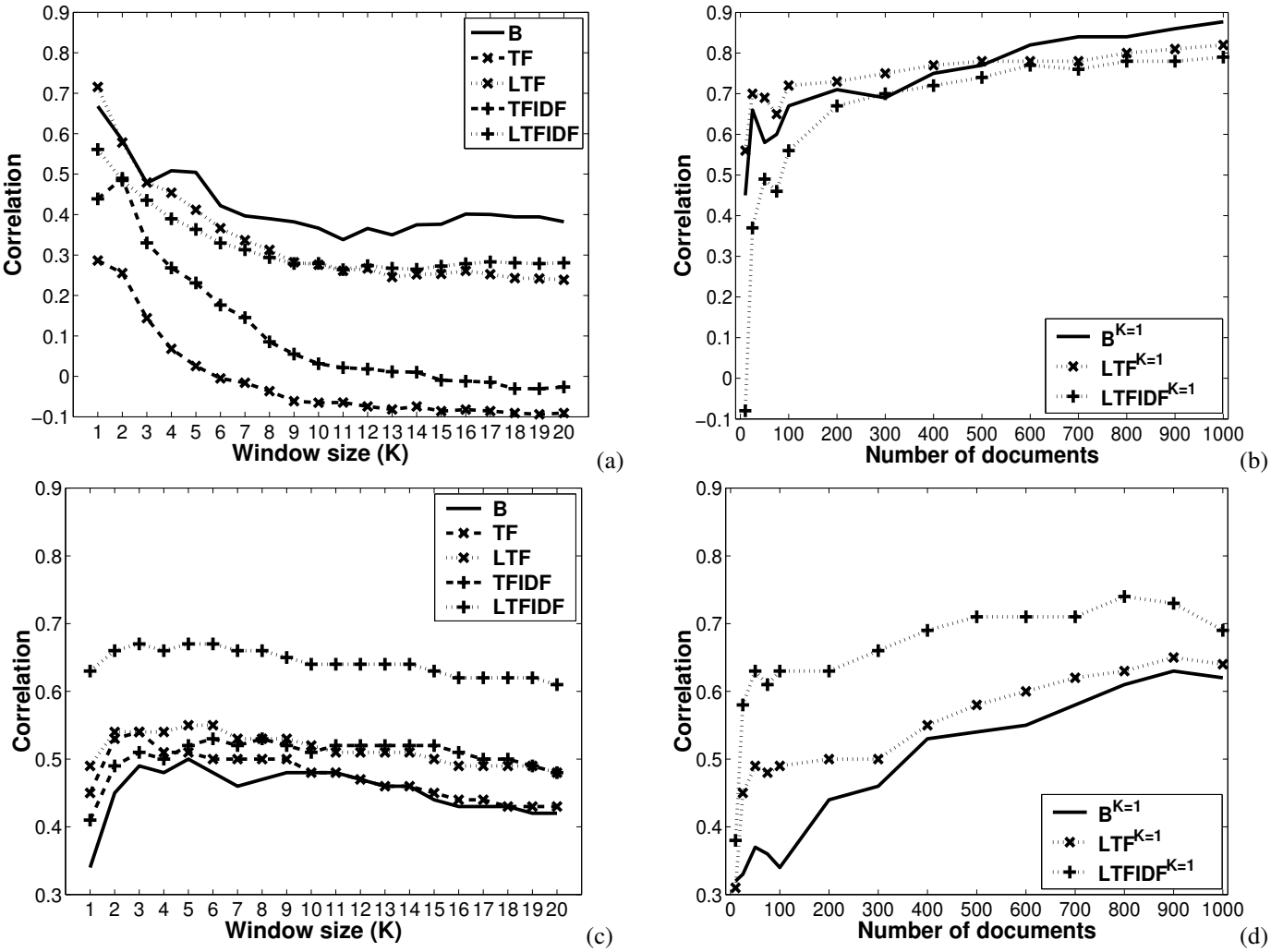

Fig. 1. Correlation scores between context-based similarity computation and human ratings for: (a),(b) the Charles-Miller dataset, and (c),(d) the MeSH dataset. Performance of the various weighting schemes as a function of context window size is shown in (a),(c) for 100 documents. Performance as a function of number of documents is shown in (b),(d) for $K=1$.

knowledge resources, i.e., supervised similarity computation algorithms. The metrics considered here, along with the main characteristics of each metric, are summarized in Table V and Table VI, for the Charles-Miller and MeSH datasets, respectively.

The Li [11], Jiang [13], X-Similarity [10], and Leacock-Chodorow [12] metrics exploit the semantic hierarchical structure of ontologies, WordNet or MeSH, to compute semantic similarity as described in Section II. The correlation scores for these metrics were taken from [10]. For the Charles-Miller dataset, correlation scores of 0.82, 0.83 and 0.74, have been reported for the Li, Jiang and X-Similarity metrics [10], respectively. Similarly for the MeSH dataset, the following correlation scores have been reported: 0.70 (Li), 0.71 (Jiang), 0.74 (Leacok-Chodorow) and 0.71 (X-Similarity) [10].

The performance of the web-based metrics is summarized as follows. For the Charles-Miller dataset, the resource-based SemSim metric proposed in [26], achieves a correlation score of 0.83 that is similar to the ontology-based methods above. The fully unsupervised Sahami [40] metric is shown to have a moderate correlation of 0.58 (results are reproduced from the implementation and evaluation in [26]). Moderate correlation scores are achieved also by the metrics that consider only the returned page counts for a query, especially for mutual information and Google. The unsupervised context-based metric using the binary weighting scheme and context window $K=1$ achieves the highest correlation (0.88) among the unsupervised metrics for 1000 documents. Note that the performance of the context-based metrics is comparable to that of the resource-based metrics for semantic similarity computation between words. In fact, the reported correlation score of 0.88 is among the highest reported in the literature for this dataset. The highest reported correlation score for the Charles-Miller dataset is equal to 0.89 [11]. The proposed algorithm exploits the shortest path length and depth between the words of interest in the WordNet hierarchy.

For the MeSH dataset, all page-count-based metrics have poor results. The best correlation (0.69) among the unsupervised metrics for 1000 documents is obtained by the context-based metric with window $S^{K=1}$ using the LTFIDF scheme. The performance of the context-based LTFIDF metric is worse but comparable to the supervised methods.

Metric	Use of ($\surd$: yes, X: no)						Need of external knowledge	Correlation
	WWW Search engine	Page counts	Snippets	Lexico-Syntactic patterns	WordNet ontology	Download documents		
Jaccard (J)	$\surd$	$\surd$	X	X	X	X	X	0.41
Dice (C)	$\surd$	$\surd$	X	X	X	X	X	0.41
Mutual information (I)	$\surd$	$\surd$	X	X	X	X	X	0.69
Google-based semantic relatedness ($G^{'}$)	$\surd$	$\surd$	X	X	X	X	X	0.66
Sahami	$\surd$	X	$\surd$	X	X	X	X	0.58
SemSim	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	X	$\surd$	0.83
Li	X	X	X	X	$\surd$	X	$\surd$	0.82
Jiang	X	X	X	X	$\surd$	X	$\surd$	0.83
X-Similarity	X	X	X	X	$\surd$	X	$\surd$	0.74
Proposed $S^{K=1}$ (B scheme over 1000 docs)	$\surd$	X	X	X	X	$\surd$	X	0.88

TABLE V

PROPERTIES AND PERFORMANCE OF SIMILARITY METRICS FOR THE CHARLES-MILLER DATASET.

Metric	Use of ($\surd$: yes, X: no)				Need of external knowledge	Correlation
	WWW Search engine	Page counts	MeSH ontology	Download documents		
Jaccard (J)	$\surd$	$\surd$	X	X	X	0.26
Dice (C)	$\surd$	$\surd$	X	X	X	0.29
Mutual information (I)	$\surd$	$\surd$	X	X	X	0.30
Google-based semantic relatedness ($G^{'}$)	$\surd$	$\surd$	X	X	X	0.41
Li	X	X	$\surd$	X	$\surd$	0.70
Jiang	X	X	$\surd$	X	$\surd$	0.71
LeacokChodorow	X	X	$\surd$	X	$\surd$	0.74
X-Similarity	X	X	$\surd$	X	$\surd$	0.71
Proposed $S^{K=1}$ (LTFIDF scheme over 1000 docs)	$\surd$	X	X	$\surd$	X	0.69

TABLE VI

PROPERTIES AND PERFORMANCE OF SIMILARITY METRICS FOR THE MESH DATASET.

VII. DISCUSSION

In this section, the evaluation results are further analyzed and explained. Specifically, we investigate the performance of the supervised and unsupervised semantic similarity computation algorithms and explain the difference in performance for words and terms. Issues such as feature selection and document selection are addressed.

A. Corpus creation and document selection

A shortcoming of web-based methods for similarity computation is that as far as the algorithm is concerned the search engine is a "black box". This is especially relevant for page-count based metrics, such as Jaccard and Google, where the number of returned hits is very much search engine dependent and changes over time. For the context-based approach, the assumption is that a search engine is a reliable provider of representative examples of language usage. Although this is a reasonable assumption, the relative ranking of documents by a search engine might affect the algorithm's performance, given than only the top ranking documents are downloaded. Another factor that affects performance is the type of web query used, as well as, the way the query is "interpreted" by the search engine.

In this work, we use a conjunction query to search for documents in which the words or terms of interest co-exist, i.e., "w_1 AND w_2". We have also noted that using a conjunction query works much better in practice than using a disjunction, i.e., "w_1 OR w_2" [28]. There are two possible explanations for the significantly better performance of corpora created using AND vs OR queries. First, co-occurrence is by itself a feature used in semantic similarity computation, e.g., page-count based similarity. Second, the created corpus is semantically more homogeneous and stylistically more consistent. Specifically, by examining occurrences of w_1 and w_2 in the same document, the topic and authoring style are the same for the context of both words. In [33], it was shown that context-based similarity metrics work much better in semantically homogeneous domains, e.g., travel reservation, than in semantically broad domains, e.g., news. Similar observations have been made for unsupervised word sense disambiguation algorithms that also employ context-based metrics, specifically, "the sense of a target word is highly consistent within any given document"[41].

As far as the relative ranking of documents by the search engine is concerned we have not observed any statistical significant effect on the performance of the context-based metrics. Specifically, we have tested the performance of the algorithm on document deciles, i.e., documents ranked 1-100, 101-200, up to 901-1000. We have found no significant effect of rank in any of these experiments, either for the word or the term task. More research is necessary (e.g., bottom ranked documents) to verify that indeed search engine ranking does

not affect context-based semantic similarity performance.

During the application of a context-based similarity metrics over the collection of downloaded documents, we assumed that the lexical features of each document have the same importance (or weight) in the similarity computation formula. In practice, however, documents are different in many ways, e.g., authoring style and author's expertise, balance between graphical and textual content. It is not uncommon in web based application to assert the "quality" of a document and exclude (or weight less) low quality documents. In our case, we experimented with a variety of "grammaticality" metrics[9] in order to establish the quality of a document. The following metrics were used to compute document grammaticality: (a) the average number of words in a paragraph, assuming that a document consisting of paragraphs of larger size is of "higher quality", (b) the fraction of document vocabulary that is included in the document compared with a Wall Street Journal corpus, i.e., selecting documents with a more formal way of writing, and (c) the perplexity of the text in the document computed using n-gram language model built from a Wall Street Journal; this feature is looking for documents with richer vocabulary and more complex syntax. The computed metrics were then used to weigh the contribution of the features extracted from each document. None of the proposed algorithms provided consistent performance improvement compared to the baseline results. This is an indication that the performance of context-based similarity metrics is not affected much by document writing style or document "quality".

B. Feature selection for word and term similarity

The evaluation results showed that co-occurrence (page-count-based metrics) can provide only rough estimates about semantic similarity. This trend is more pronounced for the specialized medical terms of the MeSH dataset. Context-based metrics achieved higher correlation scores compared to page-count-based metrics for both tasks. Overall, context seems to be the most important feature for semantic similarity computation, followed by co-occurrence. This is consistent with the observations of linguists [30]. Moreover, evaluation results showed that performance improves as the number of downloaded documents increases, which is in agreement with the statement of Schütze and Pedersen [42] "words with similar meanings will occur with similar neighbors if enough text material is available". Although it is clear that contextual similarity implies semantic similarity, the amount of context to take into account in this process, as well as, the relative weighting of the contextual features needs further investigation and is discussed next.

We have investigated various aspects of feature selection for context-based similarity metrics, namely, context window size, the use of stop-words and the relative weighting of context words. Specifically, we found that using the very immediate context (window size one) gives best performance for the Charles-Miller dataset, while a context window size between two and five words gives the best results for the MeSH dataset. In addition, stop-words were valuable features for the Charles-Miller dataset, but provided little or no information for the MeSH dataset. Finally, term frequency (TF)-based feature weighting provided good results for the Charles-Miller dataset, while term-frequency inverse document frequency (TFIDF)-based weighting provided best results for the MeSH dataset. In essence, optimal feature selection was quite different for words and terms similarity computation.

Putting together the observations from these experiments one may draw general conclusions about feature selection for context-based similarity computation between common words or between specialized terms. Note that immediate context, stop-words and very frequent contextual features typically encode mostly syntactic dependencies. Longer context, non stop-words and features with high TFIDF weights encode mostly semantic dependencies. Thus for common words, syntax seems to be the most salient feature, while for terms, semantics are more important. More research is necessary to better understand how to tune the feature selection process for specific domains, as well as, on how to better combine different types of features, e.g., fusion of syntax and semantic-based features. Note, however, that the generic feature selection and weighting algorithms presented in this work for word and term semantic similarity computation already provide good baseline performance. Also preliminary experiments indicate that the proposed algorithms perform well for other languages, e.g., Greek.

C. Similarity metrics and human cognition

As it can be seen from the experimental results the weighting scheme plays an important role for context-based semantic similarity computation. Among term frequency metrics, logarithmic scaling of context feature occurrences consistently outperforms linear scaling for both tasks. Also the binary weighting scheme seems to perform comparably to the logarithmic term frequency (LTF) weighting for both tasks. A possible explanation of these trends could lie with human perception and cognition. Note that automatically computed similarity scores are compared against scores from human annotators, thus, the logarithmic scaling of features could be explained by a nonlinearity in human perception of semantic similarity between words. If we further assume that humans acquire semantic relationships between words using contextual information (at least for common words) the binary weighting is simpler and it is more probable to approximate human cognition. More research is needed to better understand the good performance of binary and logarithmic term frequency weighting schemes. We believe that the development of computational similarity metrics can serve as an additional research tool in the field of human cognition.

A final note on the comparison between the supervised resource-based and unsupervised context-based semantic similarity computation algorithms. In this work, we have shown for the first time that unsupervised metrics achieve comparable performance to supervised resource-based ones. Comparing, however, the best results of supervised and unsupervised algorithms should be done with care, as in both cases, there is

[9]The notion of grammaticality is used here in a broad sense rather than the exact linguistic sense of conforming to a syntactic grammar.

a long list of parameters that are being "tuned" for the specific dataset, i.e., there is a danger of model overfiting. Extensive experiments on additional datasets, as well as, optimization of parameters on held-out data is required in order to draw general conclusions about the detailed performance of the algorithms. Independent of their relative performance, however, the proposed unsupervised algorithms should prove a valuable tool for populating existing ontologies with new members[10], as well as, create ontologies for new languages.

VIII. Conclusions

We presented and compared two families of unsupervised, web-based metrics for semantic similarity computation between words, namely, page-count and context-based metrics. Page-count metrics consider only hits returned by a search engine, while the proposed context-based semantic similarity algorithms download the top ranked documents returned from a web query and compute the frequency of occurrence of contextual features. The proposed algorithms be generalized and applied to other languages and do not consult any external knowledge resource. The performance of the proposed algorithms was evaluated and compared with resource-based semantic similarity metrics on the Charles-Miller dataset and the MeSH dataset of medical terms.

The page-count-based metrics produced low to mid correlation with human semantic similarity scores. Good correlation scores were obtained using the context-based metrics, achieving performance of up to 0.88 and 0.74 for the Charles-Miller and MeSH datasets, respectively. The performance achieved is comparable to that of supervised resource-based semantic similarity computation algorithms. The following conclusions can be drawn for the performance of unsupervised similarity computation algorithms: (i) context is a better feature for semantic similarity computation than co-occurrence, (ii) for the Charles-Miller dataset best results are obtained for a contextual window size of one, including stop-words as features and the LTF or binary weighting schemes, (iii) for the MeSH dataset best results are obtained for a contextual window size of two to five, excluding stop-words as features and the LTFIDF feature weighting scheme, (iv) logarithmic weighting of contextual feature outperforms linear weighting for both tasks, and, (v) performance of context-based metrics improves as the number of documents increases (with the exception of the last two data-points for the MeSH dataset). Preliminary experiments on document selection did not show significant correlation with performance. Overall, the proposed context-based algorithm provides good performance, is fully automatic, requires little computation-power and small to medium amounts of web text, and, in addition, it can be generalized and applied to other languages. In order to use the proposed algorithms in practice for ontology creation, one may use a combination of page-count and contextual metrics, i.e., use page-count metrics to identify candidates and contextual metrics to refine the similarity scores.

[10]Note that two out of the 30 noun-pairs in the Charles-Miller dataset [39] were not included in the original versions of WordNet forcing researchers to evaluate on a 28 pair subset [26].

This work is a first step towards our understanding of the potential of context-based metrics for semantic similarity computation. A variety of issues related to document selection, feature selection and feature fusion have to be further investigated. In addition, a better understanding of acquisition of semantics by humans and human cognition could lead to improved semantic similarity computation algorithms.

References

[1] S. Gauch and J. Wang, "A corpus analysis approach for automatic query expansion," in *Proceedings of the Sixth International Conference on Information and Knowledge Management*, 1997, pp. 278–284.

[2] E. Voorhees, "Query expansion using lexical-semantic relations," in *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994, pp. 61–69.

[3] R. Mihalcea and D. Moldovan, "Semantic indexing using wordnet senses," in *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval*, 2000, pp. 35–45.

[4] S. Flank, "A layered approach to nlp-based information retrieval," in *Proceedings of the COLING'98*, 1998, pp. 397–403.

[5] E. Fosler-Lussier and H.-K. Kuo, "Using semantic class information for rapid development of language models within asr dialogue systems," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, pp. 553–556.

[6] K.-C. Siu and H. Meng., "Semi-automatic acquisition of domain-specific semantic structures," in *Proc. Interspeech*, 1999, pp. 2039–2042.

[7] I. Dagan, "Similarity-based methods for word sense disambiguation," in *Proceedings of the Association for Computational Linguistics*, 1997, pp. 56–63.

[8] A. Pargellis, E. Fosler-Lussier, C.-H. Lee, A. Potamianos, and A. Tsai, "Auto-induced semantic classes," *Speech Communication*, vol. 43, no. 3, pp. 183–203, 2004.

[9] E. Iosif, A. Tegos, A. Pangos, E. Fosler-Lussier, and A. Potamianos, "Combining statistical similarity measures for automatic induction of semantic classes," in *Spoken Language Technology, 2006. SLT '06. IEEE/ACL Workshop on*, 2006, pp. 86–89.

[10] E. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, "X-similarity: Computing semantic similarity between concepts from different ontologies," *Journal of Digital Information Management*, vol. 4, no. 4, pp. 233–238, 2006.

[11] Y. Li, Z. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.

[12] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification in wordnet," in *An Electronic Lexical Database*. MIT Press, 1998, pp. 265–283.

[13] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the 10th International Conference on Research on Computational Linguistics*, 1997.

[14] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of semantic distance," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[15] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, pp. 533–536.

[16] L. Dekang, Z. Shaojun, Q. Lijuan, and Z. Ming, "Identifying synonyms among distributionally similar words," in *IJCAI*, 2003, pp. 1492–1493.

[17] T. Chklovski and P. Pantel, "Verbocean: Mining the web for fine-grained semantic verb relations," in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, 2004, pp. 33–40.

[18] E. Terra and C. L. A. Clarke, "Frequency estimates for statistical word similarity measures," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 165–172.

[19] M. Popovic and H. Ney, "Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data," in *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, 2005, pp. 212–218.

[20] S. Dumais, M. B. E. Brill, J. Lin, and A. Ng, "Web question answering: is more always better?" in *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 291–298.

[21] P. Cimano, S. Handschuh, and S. Staab, "Towards the self-annotating web," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 462–471.

[22] P. Mika, "Ontologies are us: A unified model of social networks and semantics," in *The Semantic Web ISWC 2005*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, pp. 522–536.

[23] J. Mori, T. Tsujishita, Y. Matsuo, and M. Ishizuka, "Extracting relations in social networks from the web using similarity between collective contexts," in *The Semantic Web ISWC 2006*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2006, pp. 487–500.

[24] M. Schedl, T. Pohle, P. Knees, and G. Widmer, "Assigning and visualizing music genres by web-based co-occurrence analysis," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, 2006.

[25] G. Geleijnse and J. Korst, "Tagging artists using co-occurrences on the web," in *Proceedings Third Philips Symposium on Intelligent Algorithms (SOIA 2006)*, 2006, pp. 171–182.

[26] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 757–766.

[27] J. Gracia, R. Trillo, M. Espinoza, and E. Mena, "Querying the web: A multiontology disambiguation method," in *ICWE '06: Proceedings of the 6th international conference on Web engineering*, 2006, pp. 241–248.

[28] E. Iosif and A. Potamianos, "Unsupervised semantic similarity computation using web search engines," in *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2007, pp. 381–387.

[29] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, 1990.

[30] H. Rubenstein and J. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[31] F. Jelinek, *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[32] D. Lewis, "Naive bayes at forty: The independence assumption in information retrieval," in *Proceedings of the 10th European Conference on Machine Learning ECML-98*, 1998, pp. 4–15.

[33] A. Pangos, E. Iosif, A. Potamianos, and E. Fosler-Lussier, "Combining statistical similarity measures for automatic induction of semantic classes," in *Automatic Speech Recognition and Understanding, 2005. ASRU '05. IEEE Workshop on*, 2005, pp. 278–283.

[34] R. Feldman, I. Dagan, and H. Hirsh, "Mining text using keyword distributions," *J. Intell. Inf. Syst.*, vol. 10, no. 3, pp. 281–300, 1998.

[35] P. Vitanyi, "Universal similarity," in *Proceedings of IEEE ISOC ITW2005 on Coding and Complexity*, 2005, pp. 238–243.

[36] R. Cilibrasi and P. Vitanyi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[37] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.

[38] YahooSearchAPI, http://developer.yahoo.com/search/.

[39] G. Miller and W. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1998.

[40] M. Sahami and T. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 377–386.

[41] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 189–196.

[42] H. Schütze and J. Pedersen, "Information retrieval based on word senses," in *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, 1995, pp. 161–175.

Elias Iosif was born in Limassol, Cyprus in 1980. He received the Diploma and the M.S degrees from the Dept. of Electronics and Computer Engineering, Technical University of Crete, Greece, in 2005 and 2007 respectively.

Since 2007, he is a research assistant and a Ph.D. candidate at the Dept. of ECE, Tech. Univ. of Crete.

His current research interests include natural language processing and information retrieval. He is a member of the Cyprus Scientific and Technical Chamber since 2005.

Alexandros Potamianos (M'92) received the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece in 1990. He received the M.S and Ph.D. degrees in Engineering Sciences from Harvard University, Cambridge, MA, USA in 1991 and 1995, respectively.

From 1991 to June 1993 he was a research assistant at the Harvard Robotics Lab, Harvard University. From 1993 to 1995 he was a research assistant at the Digital Signal Processing Lab at Georgia Tech. From 1995 to 1999 he was a Senior Technical Staff Member at the Speech and Image Processing Lab, AT&T Shannon Labs, Florham Park, NJ. From 1999 to 2002 he was a Technical Staff Member and Technical Supervisor at the Multimedia Communications Lab at Bell Labs, Lucent Technologies, Murray Hill, NJ. From 1999 to 2001 he was an adjunct Assistant Professor at the Department of Electrical Engineering of Columbia University, New York, NY. In the spring of 2003, he joined the Department of Electronics and Computer Engineering at the Technical University of Crete, Chania, Greece as an associate professor.

His current research interests include speech processing, analysis, synthesis and recognition, dialog and multi-modal systems, nonlinear signal processing, natural language understanding, data mining, artificial intelligence and multimodal child-computer interaction.

Prof. Potamianos has authored or co-authored over eighty papers in professional journals and conferences. He is the co-author of the paper "Creating conversational interfaces for children" that received a 2005 IEEE Signal Processing Society Best Paper Award. He holds four patents. He is a member of the IEEE Signal Processing Society since 1992 and he is currently serving his second term with the IEEE Speech Technical Committee.