A Novel Method for Word-Pair Similarity Computing

Imen Akermi[1], Rim Faiz[2]
[1]LARODEC
ISG of Tunis
Le Bardo, Tunisia
[2]LARODEC
IHEC de Carthage
2016 Carthage Présidence
Tunisia
imen.akr@hotmail.fr, Rim.Faiz@ihec.rnu.tn

ABSTRACT: *Semantic similarity between words is fundamental to various fields such as Cognitive Science, Artificial Intelligence, Natural Language Processing and Information Retrieval. According to Baeza-Yates and Neto [2] an Information Retrieval system "should provide the user with easy access to the information in which he is interested". Therefore, in this domain, relying on a robust semantic similarity measure is crucial for automatic query suggestion and expansion process. In this same context, we propose a method that uses on one hand, an online English dictionary provided by the Semantic Atlas project of the French National Centre for Scientific Research (CNRS) and on the other hand, a page counts based metric returned by a social website.*

Keywords: Web search, Semantic Similarity, User-Generated content, Information Retrieval

Received: 12 August 2012, Revised 2 October 2012, Accepted 8 October 2012

1. Introduction

Measures of the semantic similarity of words have been used for a long time in applications in Natural Language Processing and related areas, such as the automatic creation of thesauri [18], text classification [8] and information extraction and retrieval [5], [32]. Indeed, one of the main goals of these applications is to facilitate user access to relevant information.

Besides, with the development of the Semantic Web, we are witnessing the advent of more interactive media that led to a huge volume of data from blogs, discussion forums, and social websites. This great amount of information available on Web pages opens new perspectives, allowing the collaborative construction of content and development of social networks which present a collective intelligence. It is a remarkable potential that we took advantage of in our work in order to measure semantic similarity. In this same context, we propose a method that uses Web content to measure semantic similarity between a pair of words.

The rest of the document is organized as follows: Section 2 introduces the related work on semantic similarity methods. In section 3, we present our method for measuring semantic similarity between words and its evaluation in order to demonstrate its ability. Finally, we conclude with few notes and some perspectives.

2. Related work

Major Measures of the semantic similarity of words have been used for a long time in applications in Natural Language Processing and related areas, such as automatic indexing, text annotation and summarization [19], lexical selection, automatic correction of word errors in text [6], and discovering word senses directly from text [23].

A word similarity measure was also used for language modeling by grouping similar words into classes [4].

Therefore, various methods have been proposed; some methods focus on using on line dictionaries such as WordNet and Brown corpus, other methods have used corpus based metrics for measuring semantic similarity between words and recently researches tend to develop Web based methods relying on the above mentioned methods.

2.1 Dictionary-based methods

Many previous works on semantic similarity have used manually compiled taxonomies and large text corpora like WordNet.

WordNet is a large lexical database of English visualized by Patwardhan et al. [24] as a large graph or semantic network, where each node of the network represents a real world concept. The concept could be an object like a house, or an entity like a teacher, or an abstract concept like art, and so on. Every node consists of a set of words, each representing the real world concept associated with that node. Thus, each node is essentially a set of synonyms that represent the same concept. For example, the concept of a *car* may be represented by the set of words car, auto, automobile, motorcar. Such a set, in WordNet terminology, is known as a synset. A synset also has associated with it a short definition or description of the real world concept known as a gloss. The synsets and the glosses in WordNet are comparable to the content of an ordinary dictionary.

Patwardhan et al. [24] note that what sets WordNet apart is the presence of links between the synsets. Each link or edge describes a relationship between the real world concepts represented by the synsets that are linked. For example, relationships of the form "*a vehicle is a kind of conveyance*" or "*a spoke is a part of a wheel "* are defined. Other relationships include: *is opposite of, is a member of, causes, pertains to,* etc.

Most of the WordNet based works related to similarity measure, focused on computing the path length. A short path between two words reports a high similarity. Inkpen [12] invoked the example of the two words apple and orange. The path length between these two words is 3 according to figure 1.

Leacock and Chodorow [16] combined syntactic information with semantic information from WordNet in order to increase the training space in a local context classifier problem. Their measure considers only the *is-a* hierarchies of nouns in WordNet. They considered only noun words hierarchies, therefore, this measure is restricted to finding relatedness between noun concepts. The noun hierarchies are all combined into a single hierarchy by imagining a single root node that subsumes all the noun hierarchies, which ensures that a path between every path of noun synsets in this single tree exists. The semantic relatedness of two synsets is determined by calculating the shortest path between the two in the taxonomy scaled by the depth of the taxonomy. It's defined by this following formula:

$$related_{lch}(C_1, C_2) = -\log\left(\frac{shortestPath(c_1, c_2)}{2 . D}\right) \quad (1)$$

Where:

- c_1 and c_2 represent the two concepts,
- shortestpath (c_1, c_2) specifies the length of the shortest path between the two synsets c_1 and c_2,
- D is the maximum depth of the taxonomy. The value of D turns out to be 19 in WordNet.

This method assumes the size or weight of every link in the taxonomy to be equal. This is considered by Patwardhan [25] as a false assumption. He notes that lower down in the hierarchy, concepts that are a single link away are more related than such pairs higher up in the hierarchy. Some related approaches try to overcome this weakness of simple edge counting by augmenting the information present in WordNet with statistical information from large corpora.

Resnik [27] defined a similarity measure using information content. He considered that the similarity between two concepts C_1 and C2 in the taxonomy is the maximum of the information content of all concepts C that subsume both C_1 and C_2. Information content of a concept is given by the specificity or the generality of that concept. It's how the concept is related to a topic. Then the similarity between two words is calculated as the maximum of the similarity between any concepts that the words belong to. Patwardhan [25] have noticed that the Resnik measure cause an inherent ambiguity of words that poses a problem in determining the occurrence of concepts in the corpus and thus one would not be able to tell if the occurrence of the word bank in the corpus refers to the financial institution sense of the bank or to the river-bank sense.

Li et al. [17] combined structural semantic information from a lexical taxonomy and information content from a corpus in a nonlinear model. They proposed a similarity measure that uses shortest path length, depth and local density in taxonomy.

2.2 Corpus-based methods

Corpus-based methods use frequencies of co-occurrence in corpora. One can range them into the classic vector-space model (cosine, overlap coefficient, etc.) and Latent Semantic Analysis, to probabilistic methods such as information radius and mutual information. As examples of large corpora, we mention:

- The British National Corpus (BNC) (100 million words),
- The TREC data mainly newspaper text,
- The Waterloo Multitext corpus of Webpages (one terabyte),
- The LDC English Gigabyte corpus and The Web itself.

Latent Semantic Analysis (LSA) [15] is a technique in Natural Language Processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur close together in text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called Singular Value Decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by calculating the cosine coefficient. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words.

apple (sense 1)	**Orange (sense 1)**
⇒ edible fruit **⇒ procedure, green, green groceries, garden truck** **⇒ food** **⇒ solid** **⇒ substance, matter** **⇒ object, physical object** **⇒ entity**	**⇒ citrus, citrus fruit** **⇒ edible fruit** **⇒ procedure, green, green groceries, garden truck** **⇒ food** **⇒ solid** **⇒ substance, matter** **⇒ object, physical object** **⇒ entity**

Figure 1. The WordNet path length between the words apple and orange

Turney [30] defined a measure called point-wise mutual information (PMI-IR), using the page counts returned by a web search engine, to recognize synonyms. He presented an unsupervised learning algorithm for recognizing synonyms, based on statistical data acquired by querying a Web search engine. The algorithm, called PMI-IR, uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words. Pointwise Mutual Information (PMI) is a statistical approach that uses the web as data source. The similarity between two words $w1$ and $w2$ is defined by the probability of having the two words together in a corpus divided by the probability of seeing them separately. This avoids random co-occurrence when the words are frequent.

$$PMI(w_1, w_2) = \log\left(\frac{P(w1, w2)}{P(w1) \times P(w2)}\right) \quad (2)$$

Two words co-occur when they appear in the same document. If w_1 and w_2 are statistically independent, then the probability that they co-occur is given by the product $P(w_1) \times P(w_2)$. If they are not independent, and they have a tendency to co-occur, then $P(w_1, w_2)$ will be greater than $P(w_1) \times P(w_2)$.

Another similarity measure that uses second-order co-occurrences (SOC-PMI) is defined by Islam and Inkpen [13]. It looks at the words that co-occur with the two words. The method sort lists of important neighbor words of the two target words, using PMI, then it takes the shared neighbors and adds their PMI values, from the opposite list normalized by the number of neighbors. However, McDonald [20] noticed that there is a large parameter space to explore when constructing this type of measures. Collection of co-occurrence counts requires thought about parameters such as window size, ignorance/respect of sentence boundaries, number of context words/dimensions, and selection of these context words.

2.3 Web-based methods

Several approaches used the Web for measuring semantic similarity between words. A similar approach was proposed by Matsuo et al. [21]. They proposed the use of web hits for extraction of communities on the Web. By hits, we mean the number of pages returned by a search engine for a given query. They measured the association between two personal names using the Simpson coefficient, which is calculated based on the number of web hits for each individual name and their conjunction (i.e., AND query of the two names).

Sahami and Heilman [29] measured semantic similarity between two queries using snippets returned for those queries by a search engine. For each query, they collect snippets from a search engine and represent each snippet as a TF-IDF-weighted term vector. The tf-idf (term frequency–inverse document frequency)[1] is a weight often used in Information Retrieval and Text Mining.

This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. The inverse document frequency is a measure of the general importance of the term obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient:

$$idf(t) = \log\frac{|D|}{|d : t \in d|} \quad (3)$$

Where

- $|D|$: cardinality of D, or the total number of documents in the corpus
- $|d : t \in d|$: number of documents where the term t appears (i.e., $tf(t, d) \neq 0$).

If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the formula to: $1+|d : t \in d|$. Then:

$$tf\_(t, d) = tf(t, d) \times idf(t) \quad (4)$$

Once the tf-idf-weighted term vectors are obtained, they are normalized and the centroid of the set of vectors is computed. Semantic similarity between two queries is then defined as the inner product between the corresponding centroid vectors.

Chen et al. [7] developed a double-checking model using text snippets returned by a web search engine to compute semantic similarity between words. For two words *X* and *Y*, they collect snippets for each word from a web search engine. Then they count the occurrences of word *X* in the snippets for word *Y* and the occurrences of word *Y* in the snippets for word *X*. However, this method depends heavily on the search engine's ranking algorithm. Besides, although two words *X* and *Y* might be very similar, there is no reason to believe that one can find *X* in the snippets for *Y*, or vice versa.

[1] http://en.wikipedia.org/wiki/Tf*idf

Bollegala et al. [3] modified four popular co-occurrence measures; Jaccard, Simpson, Dice, and PMI (point-wise mutual information) in order to calculate similarity measures using page counts returned by a search engine for the given word pair.

The WebJaccard coefficient between words *P* and *Q*, *WebJaccard* (*P*, *Q*), is defined as:

$$WebJaccard(P,Q) = \begin{cases} 0 & \text{if } H(P \cap Q) < C \\ \dfrac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)} & Otherwise \end{cases} \quad (5)$$

- $H(P \cap Q)$: The page count for the query $(P \cap Q)$ in the search engine Google[2].
- $H(P)$: The page count for the query P.
- $H(Q)$: The page count for the query Q.

They set the WebJaccard coefficient to zero if the page count for the query $(P \cap Q)$ is less than a threshold c[3].

Similarly, they defined *WebSimpson* (*P*, *Q*), as :

$$WebSimpson(P,Q) = \begin{cases} 0 & \text{if } H(P \cap Q) < C \\ \dfrac{H(P \cap Q)}{Min(H(P), H(Q))} & Otherwise \end{cases} \quad (6)$$

They defined the WebDice coefficient as a variant of the Dice coefficient. *WebDice* (*P*, *Q*) is defined as :

$$WebDice(P,Q) = \begin{cases} 0 & \text{if } H(P \cap Q) < C \\ \dfrac{2 \times H(P \cap Q)}{H(P) + H(Q)} & Otherwise \end{cases} \quad (7)$$

They defined *WebPMI* (*P*, *Q*) as a variant form of PMI using page counts as:

$$WebPMI(P,Q) = \begin{cases} 0 & \text{if } H(P \cap Q) < C \\ \log \dfrac{\dfrac{H(P \cap Q)}{N}}{\dfrac{H(P)}{N} \times \dfrac{H(Q)}{N}} & Otherwise \end{cases} \quad (8)$$

N is the number of documents indexed by the search engine. In their work, Bollegala et al. [3] set $N = 10^{10}$, according to the number of indexed pages reported by Google.

These measures are based on the use of association ratios between words computed using the frequency of co-occurrence of words in documents. The main hypothesis of this approach is that two words are semantically related if their association ratio is high.

Bollegala et al. [3] proposed another measure which combines retrieval of information about the number of occurrences of two words (both together and individually) from a Web search engine, with retrieval of information from text snippets returned by the search engine. They automatically discover lexico-syntactic templates for semantically related and unrelated words using WordNet, and they train a support vector machine (SVM) classifier. The learned templates are used for extracting information from the text fragments returned by the search engine. This method requires extra resources for training the SVM.

Although the methods above mentioned performed well, however they have some shortcomings. In fact, one major issue behind taxonomies and corpora oriented approaches is that they might not necessarily capture similarity between proper names such

[2] http://www.google.com

[3] c = 5 in the experiments

as named entities (e.g., personal names, location names, product names) and the new uses of existing words. Furthermore, using page counts alone for Web based methods as a measure of similarity is not enough since even though two words appear in a page, they might not be related. Bollegala et al. [3] gave a perfect exemplary case for this matter; page counts for the word *apple* contains page counts for apple as a *fruit* and apple as a *company*. Moreover, given the scale and noise in the Web, some words might occur arbitrarily, i.e. by random chance, on some pages. For those reasons, methods relying exclusively on the Web are unreliable when measuring semantic similarity.

3. A new method for measuring semantic similarity between words

With the development of the Semantic Web, it became interesting to exploit web content in order to measure the semantic similarity. In this same context, we propose a method that uses Web content to measure semantic similarity between words.

3.1 Proposed method

As described in figure 2, our method for measuring semantic similarity between words uses, on one hand, an on line English dictionary provided by the Semantic Atlas project (SA) [26] and on the other hand, page counts returned by a social website whose content is generated by users.

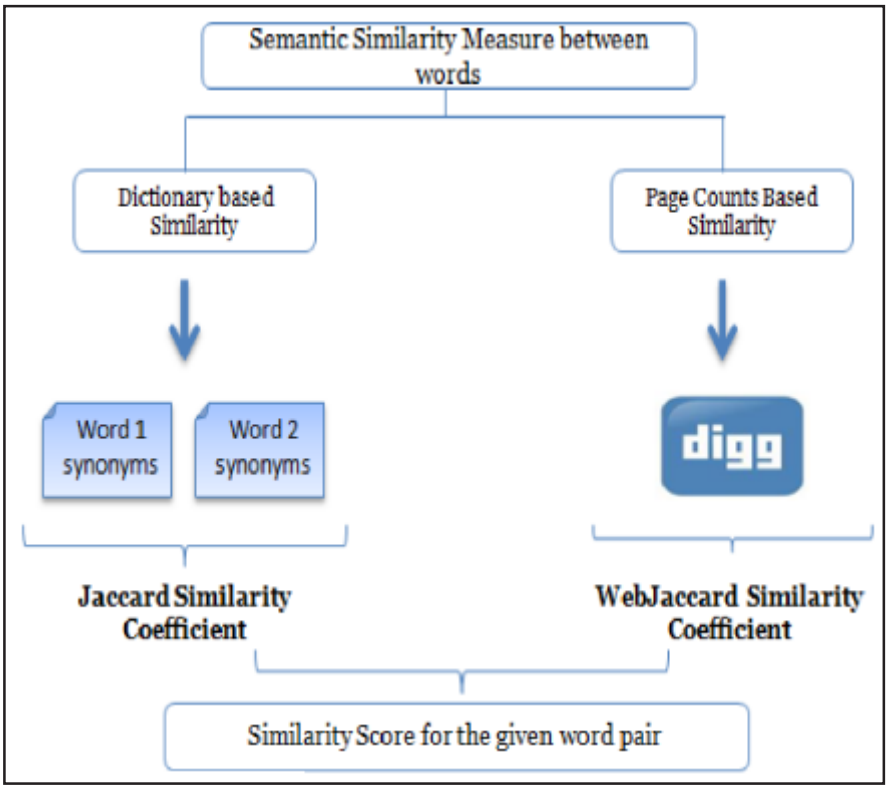

Figure 2. The proposed method for measuring semantic similarity between a given pair of words

It consists in three phases:

1. The calculation of the similarity (SD) between two words based on the online dictionary provided by the Semantic Atlas project.

2. The calculation of the similarity (SP) between two words based on the page counts returned by the social Website Digg.com[4].

3. Integration of the two similarity measures SD and SP.

[4] http://www.digg.com

3.1.1 Phase 1: The Semantic Atlas based similarity measure

In this phase, we extract synonyms for each word from the on line English dictionary provided by the Semantic Atlas (SA) project [26] of the French National Center for Scientific Research (CNRS).

The SA is composed of several dictionaries and thesauri (including the Roget's thesaurus), offering thus a wide range of senses for a given word. In fact, the SA is used for the automatic treatment of polysemy and semantic disambiguation [31].

The SA is currently available for French and English versions. It can be consulted on line via a flexible interface allowing for interactive navigation[5]. The current online version allows users to set the search type for English words as (1) *standard* (narrow synonymy) or (2) *enriched* (broad synonymy).

Once the two sets of synonyms for each word are collected, we calculate the degree of similarity, which we call $S(w_1, w_2)$, between them, using the Jaccard coefficient:

$$S(w_1, w_2) = \frac{m_c}{m_{w1} + m_{w2} - m_c} \tag{9}$$

Where

m_c: The number of common words between the two synonym sets.

m_{w1}: The number of words contained in the w_1 synonym set.

m_{w2}: The number of words contained in the w_2 synonym set.

3.1.2 Phase 2: The page counts similarity measure

In this phase, we calculate the degree of similarity between the two words w_1 and w_2 using the WebJaccard coefficient [3] which has as parameters the number of pages returned by the social Website Digg.com for queries w_1, w_2 and (w_1, w_2).

Barlow [1] identifies Digg.com as one of the most popular aggregators of articles published on the Web. We use this social site to calculate the number of pages for a given query. Therefore, once the page counts for queries w_1, w_2 and (w_1, w_2) are obtained, we calculate the WebJaccard coefficient for the given pair of words:

$$WebJaccard(w_1, w_2) = \frac{H(w_1 \cap w_2)}{H(w_1) + H(w_2) - H(w_1 \cap w_2)} \tag{10}$$

Where:

$H(w_1 \cap w_2)$: The page counts for the query (w_1, w_2).

$H(w_1)$: The page counts for the query w_1.

$H(w_2)$: The page counts for the query w_2.

Bollegala et al. [3] used the web search engine Google[6] to get page counts for a given query. However, the Google API only allows 100 automatic queries per day, and if we want to exceed the 100 requests, a charge must be paid. Fukazawaa and Ota [9] faced the same problem in their work. We did not encounter this issue with the API provided by digg.com.

3.1.3 Phase 3: The overall similarity measure

In this last phase, we incorporate both measures previously calculated by the following formula:

[5] http://dico.isc.cnrs.fr: This site is the most consulted address of the French National Center for Scientific Research's domain (CNRS), one of the major research bodies in France

[6] http://www.google.com

$$Sim_{FA}(w_1, w_2) = \alpha \times S(w_1, w_2) + (1- \infty) \times WebJaccard(w_1, w_2) \quad (11)$$

$\alpha \in [0, 1]$.

First experiments on Miller-Charles [22] and on Rubenstein-Goodenough's [28] datasets have shown that our measure performs better with $\alpha = 0,6$.

3.2 Evaluation

3.2.1 Corpus based evaluation

We evaluated the proposed method against Miller-Charles [22] and Rubenstein-Goodenough's [28] datasets. These two datasets are considered as a reliable benchmark for evaluating semantic similarity measures. The word pairs are rated on a scale from 0 (no similarity) to 4 (perfect synonymy).

3.2.1.1 The Miller-Charles's dataset
Miller and Charles chose 30 pairs from the original Rubenstein-Goodenough's 65 word pairs, taking 10 from the "*high level* (*human score is between* 3 *and* 4), 10 f*rom the intermediate level* (*human score is between* 1 *and* 3), *and* 10 *from the low level (human score is from* 0 *to* 1) *of semantic similarity*", and then obtained similarity judgments from 38 subjects, on those 30 pairs. Each subject was tested individually. The subjects were told to judge similarity of meaning. As examples of degrees of synonymy, the pairs *gem-jewel*, *bird-cock*, and *autograph-shore*, which were rated high, intermediate, and low respectively by Rubenstein and Goodenough [28], were shown to each subject. A subject was then presented with the two sheets of paper on which 30 noun pairs appeared and was instructed to examine each pair closely and then to rate it on a 5-point scale from 0 to 4, where 0 represents no similarity of meaning and 4 perfect synonymy. The ordering of the pairs was randomly determined for each subject. The subjects were free to rate and re-rate the pairs for as long as they chose.

The sets of ratings are in good correspondence with the Rubenstein-Goodenough's: the Pearson product-moment correlation coefficient is 0.97, significant at the 0.01 level. People are not only able to agree reasonably well about the semantic distances between concepts, but their average estimates remain remarkably stable over more than 25 years [22].

In table 1, we compare the results of our measure against Miller-Charles dataset with:

- Chen's Co-occurrence Double Checking (CODC) measure [7]
- Sahami and Heilman's [29]
- Hirst and St-Onge's [10]
- Leacock and Chodorow's [16]
- Lin's [18]
- Resnik's [27]
- Bollegala et al.'s [3]

According to table 1, our proposed measure Sim_{FA} earns the highest correlation of 0,836.

2.3.1.2 The Rubenstein-Goodenough's dataset
Rubenstein and Goodenough [28] obtained "*synonymy judgements*" from 51 human subjects on 65 pairs of words. The pairs ranged from "*highly synonymous*" to "*semantically unrelated* ", and the subjects were asked to rate them, on the scale of 0.0 to 4.0, according to their "*similarity of meaning*".

We evaluate, in Table 2, our method against Rubenstein and Goodenough's dataset with the following measures:

- Hirst and St-Onge's [10]
- Jiang and Conrath's [14]
- Leacock and Chodorow's [16]

Method	Correlation
Hirst and St Onge	0,744
Leacock and chodorow	0,816
Lin	0,829
Resnik	0,774
Sahami et al.	0,579
Chen - CODC	0,693
Bollegala et al.	0,834
Proposed Sim_{FA}	**0,836**

Table 1. The Sim_{FA} similarity measure compared to baselines on Miller-Charles' dataset

- Lin's [180]
- Resnik's [27]
- Li et al. [17]
- Bollegala et al. [3]

The first is claimed as a measure of semantic relatedness because it uses all noun relations in WordNet; the others are claimed only as measures of similarity because they use only the hyponymy relation. These measures were implemented by Budanitsky and Hirst [6], where the Brown Corpus was used as the basis for the frequency counts needed in the information-based approaches.

Method	Correlation
Hirst and St Onge	0,786
Leacock and chodorow	0,838
Lin	0,819
Resnik	0,779
Li et al.	0,891
Jiang and Conrath	0,781
Bollegala et al.	0,812
Proposed Sim_{FA}	**0,866**

Table 2. The Sim_{FA} similarity measure compared to baselines on Rubenstein-Goodenough's dataset

As summarized in table 2, our proposed measure evaluated against Rubenstein-Goodenough's dataset gives a correlation coefficient of 0,866 which we can consider as promising.

In fact, our measure outperforms simple WordNet-based approaches such as Edge counting and Information Content measures and it is comparable with the other methods. Our proposed method does not require hierarchical taxonomy of concepts or sense-tagged definitions of words, unlike the WordNet based methods.

Therefore, it can be used to calculate semantic similarity between named entities, which are not fully covered by WordNet or other manually compiled thesauri.

2.3.1.3 Discussion

In order to calculate the page counts for a given query, Bollegala et al. [3] used the web search engine Google. However, the Google API only allows 100 automatic queries per day, and if we want to exceed the 100 requests, a charge must be paid. We did

not encounter this issue with the API provided by Digg.com.

However, we went further in our experiments by evaluating our method against the two datasets mentioned earlier, using the page counts returned by the web search engine Google instead of those returned by the social website Digg.com, so that we could be sure that the results returned by Digg.com were as accurate as those returned by Google.

DataSet / **Page counts returned by**	Miller-Charles	Rubenstein-Goodenough
Google.com	**0,828**	**0,865**
Digg.com	**0,836**	**0,866**

Table 3.The SimFA measure evaluated with Google.com and Digg.com

According to table 3, the *SimFA* measure performs slightly better when we use, as basis for page counts, the social website Digg.com. This emphasizes the fact that we can use the social web site Digg.com instead of the search engine Google, in order to get automatically and with no charges, page counts for a given dataset exceeding the Google limit.

3.2.2 Task based evaluation

Inkpen [12] notes that we can't only rely on the correlation with the human judges in order to measure the performance of a similarity metric. We can't deny that it is a recommended evaluation step; however, it is not sufficient because it can be done only on a small set of noun pairs. Therefore, a task-based evaluation section is recommended in order to fully cover the performance of a word similarity measure.

In order to determine the ability of our similarity metric to provide improvement in tasks such as classification, we classified 113 terms expressing negative and positive opinions used by Elkhlifi et al. [8] for extracting and classifying opinions based on the semantic similarities between words. Our measure correctly classified 100 terms on a total of 113, which gives a Percentage of Perfect Classification, denoted PCC of 88,5% which we can consider as promising:

$$PCC = \frac{\text{Number of well classified terms}}{\text{Total Number of terms}} \times 100 \qquad (12)$$

4. Conclusion

Semantic similarity measures have been the central concern of taxonomists of the previous century, and experiences now-a-days a major revival of interest inherent in the evolution of new technologies of Natural Language Processing.

The increasing complexity of data requires the development of measures able to keep a semantic relevance with respect to the application domain. In fact, semantic similarity is fundamental to various fields such as Cognitive Science, Artificial Intelligence, Natural Language Processing and Information Retrieval.

Several studies on Natural Language Processing were also motivated by semantic similarity measures, such as the work of Hirst and Budanitsky [11]. They investigate the usefulness of the semantic similarity in the problem of spelling correction, where actual spelling errors are detected and corrected automatically. This accentuates the importance of relying on a reliable and robust similarity measure.

Therefore, various methods have been proposed; Corpus-based methods using online dictionaries and Web-based metrics. In this paper, we introduced a new similarity measure between words combining on one hand, the use of an online English dictionary provided by the Semantic Atlas project of the French National Centre for Scientific Research (CNRS) and on the other hand, page counts returned by the social website Digg.com.

Experimental results have shown that our proposed method is promising.

There are several lines of future work that our proposed measure lays the foundation for. We will incorporate this measure into

other similarity-based applications to determine its ability to provide improvement in tasks such as clustering of text. Besides, we will take advantage of several other characteristics of the social website Digg.com in order to measure semantic similarity.

References

[1] Barlow, A. (2008). Blogging America: The New Public Sphere. Praeger.

[2] Baeza-Yates, R., Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley.

[3] Bollegala, D., Matsuo, Y., Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. *In:* Proc. of International Conference on the World Wide Web (WWW 2007), p. 757–766.

[4] Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., Mercer, R. L. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18 (4) 467–479.

[5] Buckley, C., Salton, G., Allan, J., Singhal, A. (1994). Automatic query expansion using smart: Trec 3. *In:* Proc. of 3rd Text Retreival Conference, p. 69-80.

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

[7] Chen, H., Lin, M., Wei, Y. (2006). Novel association measures using web search with double checking. *In:* Proc. of the COLING/ACL, p. 1009–1016.

[8] Elkhlifi, A., Bouchlaghem, R., Faiz, R (2011). Opinion Extraction and Classification Based on Semantic Similarities. *In*: *P*roc. of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference (FLAIRS'11), AAAI Press.

[9] Fukazawa, Y., Ota, J. (2010). Automatic modeling of user's real world activities from the web for semantic ir. *In*: Proc. of the 3rd International Semantic Search Workshop, SEMSEARCH '10, 5,1–5:9, New York, NY, USA, ACM.

[10] Hirst, G., St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *In Christiane Fellbaum editor, WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, 305–332.

[11] Hirst, G., Budanitsky, A. (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11, 87–111.

[12] Inkpen. D. (2007). Semantic similarity knowledge and its applications. Studia Universitatis BabesBolyai Informatica, LII (1)11–22.

[13] Islam, A., Inkpen, D. (2006). Second order co-occurrence pmi for determining the semantic similarity of words. *In*: Proc. of the International Conference on Language Resources and Evaluation, p. 1033–1038.

[14] Jiang, J. J., Conrath, D.W (1998). Semantic similarity based on corpus statistics and lexical taxonomy. *In*: Proc: of the International Conference on Research in Computational Linguistics ROCLING X.

[15] Landauer, T. K., Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, *Induction, and Representation of Knowledge. Psychological Review,* 104 (2) 211–240.

[16] Leacock, C., Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *In*: Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, 265–283.

[17] Li, Y., Bandar, Z., McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* 15 (4) 871-882.

[18] Lin, D. (1998). An information-theoretic definition of similarity. *In:* Proc. of the 15th ICML, p. 296–304.

[19] Lin, C., Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *In*: Proc. of the Conference of 71 Bibliography the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL '03, 1, p. 71–78, Stroudsburg, PA, USA.

[20] McDonald, S (1997). Exploring the validity of corpus-derived measures of semantic similarity. *In*: Proc. of the 9th Annual CCS/HCRC Postgraduate Conference, University of Edinburgh.

[21] Matsuo, Y., Sakaki, T., Uchiyama, K., Ishizuka, M. (2006). Graph-based word clustering using web search engine. *In:* Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP'06.

[22] Miller, G.A, Charles, W.G (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 (1).

[23] Pantel, P., Lin, D. (2002). Discovering word senses from text. *In*: Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, p. 613–619, ACM, New York, NY, USA.

[24] Patwardhan, S., Banerjee, S., Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. *In*: Proc. of the 4th international conference on Computational linguistics and intelligent text processing, CICLing'03, p. 241–257, Berlin, Heidelberg, Springer-Verlag.

[25] Patwardhan, S. (2003). Incorporating dictionary and corpus information into a vector measure of semantic relatedness. PhD thesis, University of Minnesota.

[26] Ploux, S., Boussidan, A., Ji, H. (2010). The Semantic Atlas: and interactive Model of Lexical Representation. *In:* Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta.

[27] Resnik, P. (1995). Using information content to evaluate semantic similarity in taxonomy. *In:* Proc. of 14th International Joint Conference on Artificial Intelligence (IJCAI).

[28] Rubenstein, H., Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM 8*, p. 627-633.

[29] Sahami, M., Heilman, T. (2006). A web-based kernel function for measuring the similarity of short text snippets. *In:* Proc. of 15th International World Wide Web Conference (WWW'06).

[30] Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. *In:* Proc. of ECML, p. 491–502.

[31] Venant, F. (2007). Utiliser des classes de sélection distributionnelle pour désambiguïser les adjectifs, *In:* Proc. of TALN, Toulouse, France.

[32] Xu, J., Croft, B. (2000). Improving the effectiveness of information retrieval, *ACM Transactions on Information Systems*, 18 (1) 79-112.

Authors Biographies

Imen Akermi received the Bachelor of Science degree (B.S.) in Computer Science applied to Management in June 2010 from the High Institute of Management of Tunis. She received the Master of Science (M.S.) in Computer Science applied to Management in 2012. She is now a PhD Student in Computer Science between University of Tunis and Paul Sabatier University, France. Her research is focused on Natural Language Processing, Information Retrieval, Text Mining, and Semantic Web.

Dr. Rim Faiz She obtained her Ph.D. in Computer Science from the University of Paris-Dauphine, in France. She is currently a Professor in Computer Science at the Institute of High Business Study (IHEC), University of Carthage, in Tunisia. Her research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Information Retrieval, Text Mining, and Semantic Web. She has several publications in international conferences such as IEEE, ACM and AAAI. She is member of scientific and organization committees of several international conferences. Dr. Faiz is also the responsible of the Professional Master "E-Commerce" and the Research Master "Business Intelligence applied to the Management" at IHEC, University of Carthage.