Comparisons of Relatedness Measures Through a Word Sense Disambiguation Task

Didier Schwab, Andon Tchechmedjiev, Jérôme Goulian and Gilles Sérasset

Abstract Michael Zock's work has focussed these last years on finding the appropriate and most adequate word when writing or speaking. The semantic relatedness between words can play an important role in this context. Previous studies have pointed out three kinds of approaches for their evaluation: a theoretical examination of the desirability (or not) of certain mathematical properties, for example in mathematically defined measures: distances, similarities, scores, …; a comparison with human judgement or an evaluation through NLP applications. In this article, we present a novel approach to analyse the semantic relatedness between words that is based on the relevance of semantic relatedness measures on the global level of a word sense disambiguation task. More specifically, for a given selection of senses of a text, a global similarity for the sense selection can be computed, by combining the pairwise similarities through a particular function (sum for example) between all the selected senses. This global similarity value can be matched to other possible values pertaining to the selection, for example the F1 measure resulting from the evaluation with a gold standard reference annotation. We use several classical local semantic similarity measures as well as measures built by our team and study the correlation of the global score compared to the F1 values of a gold standard. Thus, we are able to locate the typical output of an

D. Schwab (✉) · A. Tchechmedjiev · J. Goulian · G. Sérasset
Université de Grenoble Alpes, LIG-GETALP, Grenoble, France
e-mail: Didier.Schwab@imag.fr
URL: http://getalp.imag.fr/WSD/

A. Tchechmedjiev
e-mail: Andon.Tchechmedjiev@imag.fr
URL: http://getalp.imag.fr/WSD/

J. Goulian
e-mail: Jerome.Goulian@imag.fr
URL: http://getalp.imag.fr/WSD/

G. Sérasset
e-mail: Gilles.Serasset@imag.fr
URL: http://getalp.imag.fr/WSD/

N. Gala et al. (eds.), *Language Production, Cognition, and the Lexicon*,
Text, Speech and Language Technology 48, DOI 10.1007/978-3-319-08043-7_13

algorithm compared to an exhaustive evaluation, and thus to optimise the measures and the sense selection process in general.

Keywords Semantic relatedness · Word sense disambiguation · Semantic similarity measures · Evaluation of semantic similarity measures · Best atteignable score · Correlation global score/F1 measure · Lesk measures · Gloss overlap measures · Tversky's similarity measure · Gloss vector measure

1 Introduction

Michael Zock's work has focussed these last years on finding the appropriate and most adequate word when writing or speaking (Zock et al. 2010; Zock and Schwab 2011). The semantic relatedness between words can play an important role in this context. Previous studies have pointed out three kinds of approaches for their evaluation: a theoretical examination of the desirability (or not) of certain mathematical properties, for example in mathematically defined measures: distances, similarities, scores, …; a comparison with human judgement or an evaluation through NLP applications.

In this article, we present a novel approach to analyse the semantic relatedness between words that is based on the relevance of semantic relatedness measures on the global level of a word sense disambiguation task. More specifically, for a given selection of senses of a text, a global similarity for the sense selection can be computed, by combining the pairwise similarities through a particular function (sum for example) between all the selected senses. This global similarity value can be matched to other possible values pertaining to the selection, for example the F1 measure resulting from the evaluation with a gold standard reference annotation.

We use several classical local semantic similarity measures as well as measures built by our team and study the correlation of the global score compared to the F1 values of a gold standard. Thus, we are able to locate the typical output of an algorithm compared to an exhaustive evaluation, and thus to optimise the measures and the sense selection process in general.

In this article, we first present the notion of similarity measures and we give some examples of measures that can be used on words of any part of speech. Secondly, we present the evaluation of similarity measures in the state of the art before introducing our proposition of a new evaluation method. To that end, we first present Word Sense Disambiguation (WSD) and, the ways the task can be evaluated and then we present our own method by introducing two metrics: the best atteignable score and the correlation between the global score and the F1 measure. We test it on five semantic similarity measures: two implementations of the Lesk and extended Lesk measures (one implementation from our team and one implementation from Pedersen's WordNet similarity library) and Pedersen's implementation of the gloss vector measure.

2 Similarity Measures and Their Evaluation

For most natural language processing methods and applications, there is a need to determine lexico-semantic relatedness between word senses, words or text segments. The goal is mainly to determine whether two words or text segments have some closeness in their meanings. We focus in this article on resource-based measures of semantic relatedness that have been proposed for use in natural language applications. In this context, four principal categories of semantic relatedness measures can be distinguished: feature based measures, taxonomic path length measures, information-based measures and hybrid measures. For a complete state of the art, the reader can refer for instance to Budanitsky and Hirst (2006), Cramer et al. (2010), Pedersen et al. (2005) or Navigli (2009). We briefly present the features based measures that we aim at evaluating (Lesk, Extended Lesk and Gloss Vector Measures).

2.1 Features Based Measures

Semantic relatedness measures have first been studied in the context of cognitive psychology and involve the consideration of features that characterize (positively or negatively) the similarity of two objects.

2.1.1 Tversky's Similarity Measure

Tversky (1977) first proposed an approach based on the overlap of features between two objects. The similarity between two objects is expressed as the number of pondered common properties minus the pondered specific properties of each object. The proposed model is therefore non symmetric (Fig. 1).

Formally, reprising the notations of Pirrò and Euzenat (Pirró and Euzenat 2010) where $\Psi(s)$ is the feature's set of a sense s, the Tversky's similarity can be expressed by:

$$sim_{tvr}(s_1, s_2) = \theta F(\Psi(s_1) \cap \Psi(s_2)) - \alpha F(\Psi(s_1) \setminus \Psi(s_2)) - \beta F(\Psi(s_2) \setminus \Psi(s_1))$$

where F is a function that expresses feature relevance, where $\setminus$ denotes the set difference operator and where θ, α and β respectively denote the relative importance between senses similarity, the dissimilarities between s_1 and s_2, and the dissimilarities between s_2 and s_1.

This measure can be normalized (with $\theta = 1$):

$$sim_{tvtr}(s_1, s_2) = \frac{F(\Psi(s_1) \cap \Psi(s_2))}{F(\Psi(s_1) \cap \Psi(s_2)) + \alpha F(\Psi(s_1) \setminus \Psi(s_2)) + \beta F(\Psi(s_2) \setminus \Psi(s_1))}$$

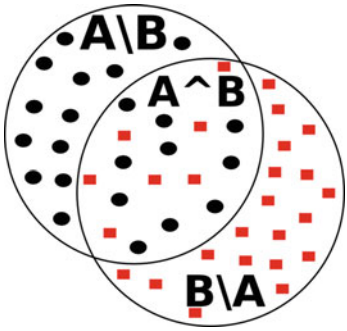

Fig. 1 Contrast between two objects

As mentioned by Pirrò and Euzenat (2010), depending on the values of α and β the Tversky index becomes one of several feature overlap similarity measures. If $\alpha = \beta = 0$, only the common features between the two senses are taken into account. If $\alpha > \beta$ or $\alpha < \beta$ we focus asymmetrically on the similarity of s_1 with s_2 or of s_2 with s_1. If $\alpha = \beta \neq 0$ the mutual similarity between s_1 and s_2 is considered. When $\alpha = \beta = 1$ Tversky's similarity measure is then equal to Tanimoto's index (Rogers and Tanimoto 1960). When $\alpha = \beta = 0.5$ the similarity is equivalent to the Dice coefficient (Dice 1945).

2.1.2 The Lesk Similarity Measure

Lesk proposed more than 25 years ago, a very simple algorithm for lexical disambiguation that evaluates the similarity between two senses as the number of common words (space separated tokens) in the definition of the senses in a dictionary (Lesk 1986) . In the original version, neither word order in the definitions (bag-of-words approach), nor any syntactic or morphological informations are taken into account. In this context, it appears that such a method can be seen as a particular case of Tversky's similarity with $\alpha = \beta = 0$ and where $\Psi(s) = D(s)$ is the set of words in the definition of s. We have:

$$sim_{lesk}(s_1, s_2) = |D(s_1) \cap D(s_2)|$$

This similarity measure is thus very simple to evaluate and only requires a dictionary and no training. The original Lesk algorithm evaluated similarity exhaustively between all senses of all words in the context. According to Navigli (2009), there are variants that select the best sense by computing the relatedness between the definition of the sense and the words in the surrounding context (with a fixed window size), rather than computing the score of all sense combinations. The similarity thus corresponds to the overlap of the sense's definition and a bag of words that contains all the words of the definitions of the context words: $Lesk_{var} = |context(w) \cap D(s_{w_n})|$. As pointed out by Navigli (2009), one important

problem of Lesk's similarity is that it is very sensitive to the words that are present in the definition; if important words are missing in the definitions used, the quality of the results will be worse. Moreover, if the definitions are too concise (as it is often the case) it is difficult to obtain fine distinctions between the similarity scores. However, many improved measures, derived from Lesk, have been proposed, as detailed in the next section.

2.1.3 Extended Lesk Measures

Wilks and Stevenson (1998) have proposed to give weights to each word of the definition, depending on the length of the definition, in order to give the same importance to all definitions instead of systematically favoring the longest definitions.

More recently Banerjee and Pedersen (2002) have proposed an extended Lesk measure that considers not only the definition of a sense but also the definitions of related sense through taxinomical links in WordNet. To calculate the overlap between two senses, they propose to consider the overlap between the two definitions of the senses but also between the definitions from different relationships: hyperonymy, hyponymy, meronymy, holonymy and troponymy but also the relations *attribute*, *similar-to*, *also-see*.

To ensure that the measure remains symmetric, the overlap is evaluated on pairs of similar relations in retaining a pair relations (R_1, R_2) only if the reverse pair (R_2, R_1) is present. This produces a set *RELPAIRS*. In addition, the overlap between the two definitions A and B, is calculated as the sum of the squares of the lengths of all substrings of words from A to B, which is expressed with the $\cap$ operator. We have:

$$Lesk_{extended}(s_1, s_2) = \sum_{\forall (R_1, R_2) \in RELPAIRS^2} (|D(R_1(s_1)) \cap D(R_2(s_2))|)$$

2.1.4 Gloss Vector Measure

Similarly, relations in WordNet are used by Patwardhan and Pedersen (2006) to augment glosses for their Gloss Vector measure. This measure combines the structure and content of WordNet with co-occurence information derived from raw text. The idea is based on textual context vectors (second order co-occurrence vectors) and was created by Schutze (1998) for the purpose of Word Sense Discrimination. Word senses are represented by second-order co-occurrence vectors of their WordNet definitions. The relatedness of two senses is then computed as the cosine distance of their representative gloss vectors. This measure allows comparisons between any two concepts without regard to their parts of speech.

2.2 Evaluation of Relatedness Measures

It is commonly accepted that there are three ways to evaluate a semantic similarity measure:

- through a theoretical point of view with its mathematical properties as these scores may be similarities (in the mathematical sense) and therefore have a value between 0 and 1, distances—and therefore satisfy the axioms of reflexivity, symmetry and triangle inequality—and so on;
- through the comparison to human judgement if it is possible to collect a large set of reliable subjects-independent judgments;
- through the performance of the measures in the context of a particular application.

In this article, we will use the third method to compare different measures in the framework of a Word Sense Disambiguation (WSD) task. We focus on three features-based measures that use WordNet: Lesk (Lesk 1986), Extended Lesk (Banerjee 2002), and Gloss Vector measure (Patwardhan and Pedersen 2006) that have been presented Sect. 2.1.

3 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is an essential task in Natural Language Processing applications, as it deals with the resolution of lexico-semantic ambiguities in natural language texts. Let us first make a general introduction of what constitutes a WSD system.

3.1 Generalities

The Word Sense Disambiguation process can be divided in three main steps:

1. build or select raw lexical material(s) (dictionaries, lexical databases, unannotated corpora, sense annotated corpora, …);
2. build an elaborate resource (a computational representation of a inventory of possible word senses);
3. use this resource to lexically disambiguate a text.

Lexical resources thus constitute a crucial element of Word Sense Disambiguation algorithms. The principle behind such algorithms is to exploit one or more resources and extract as many useful and meaningful features as possible, in order to disambiguate a text. Naturally, if no features can be selected and extracted from a given resource, the algorithm will have nothing to work on, thus limiting the

usefulness of that particular resource. In this manner, feature selection and extraction are a key process to the success of any WSD algorithm and are strongly dependent on the type and quality of the resources exploited.

3.2 Evaluation of Word Sense Disambiguation

We will first present the principle that governs the evaluation of WSD algorithms, followed by a description of how gold standards are built and what evaluation metrics are customarily used.

3.2.1 Principle

There are two means of evaluating Word Sense Disambiguation algorithms:

- In vivo evaluation, where WSD systems are evaluated through their contributions to the overall performance of a particular NLP application. It is the most natural evaluation method, but also the harder to set up.
- In vitro evaluation where the WSD task is defined independently of any particular application. In this case, systems are evaluated using specifically constructed benchmarks.

In this article, we are more particularly going to focus on the in vitro approach

3.2.2 Gold Standard

In vitro evaluation uses a reference sense-annotated corpus. In WSD, several sense-annotated corpora are commonly used:

- The Defense Science Organization corpus provided by Ng and Lee (1996), is a non-freely available sense-annotated English corpus. 192,800 word occurrences were manually tagged with WordNet synsets. The annotation of this corpus covers 121 nouns (113,000 occurrences) and 70 verbs (79,800 occurrences) that are both the most frequent and the most ambiguous. The authors claim that their coverage corresponds to 20 % of verb and noun occurrences in English texts.
- SemCor (Miller et al. 1993) is a subset of the Brown Corpus (1961). Out of 700,000 words, almost 230,000 words are manually tagged with Wordnet synsets, over a span of 352 texts. In 186 of the texts, 192,639 (all nouns, verbs, adjectives, and adverbs) are annotated, while on the remaining 166, only 41,497 verbs are annotated.
- BabelCor (Navigli 2012) is certainly the most recent annotated corpus as it was released in July 2012. It is a corpus annotated with Babel synsets. It is constituted of two parts. The first is built from SemCor, where each WordNet

synset is simply mapped to the corresponding BabelNet synsets and the other is built from Wikipedia where each hyperlink is similarly mapped to the corresponding Babel synsets.
- Corpora from evaluation campaigns: Since 1998, there have been several campaigns (SemEval–SensEval) to evaluate Word Sense Disambiguation over several languages. Most of them have been English evaluation tasks, but there have also been Japanese, Spanish and Chinese tasks. It is uncommon for WSD evaluation corpora to go beyond 5,000 tagged words.

The three first corpora are commonly used in WSD to build supervised WSD systems (WSD systems based on machine learning principles), the last ones for evaluation. We choose one text of the Semeval 2007 corpus to illustrate the method introduced here.

3.2.3 Metrics

In WSD tasks, four standard metrics are traditionally used to evaluate the quality of the solutions provided (Navigli 2009):

The first metric is Coverage (C) and is defined as the number of answers provided over the number of expected answers, in other words it represents how much of the text has been disambiguated.

The second metric is Precision (P) and is defined as the number of correct answers provided over the total number of answers provided.

The third is Recall (R) and is defined as the number of correct answers provided over the total number of answers expected to be provided.

The last metric is the F1 measure represents the "weighted harmonic mean of Precision and Recall" and combines that P and R in a single measure. It is defined as $F_1 = \frac{2 \cdot P \cdot R}{P+R}$.

3.3 Similarity-Based Word Sense Disambiguation

3.3.1 Principle

Similarity-based methods for WSD rest on two algorithms: a local algorithm and a global algorithm. Local algorithms aim at providing a score based on the proximity of the semantic content of two compared linguistic items (usually words or word senses). Similarity measures have been described in Sect. 2. For WSD, these measures are used locally between two senses, and are then applied on the global level. A global algorithm is a method that allows to extend a local algorithm to an entire text in order to infer the appropriate sense for each word. The most direct algorithm is the exhaustive (brute force) method, used for example by Banerjee and Pedersen (2002). The combinations of all the senses of the words in a given

context (word window or text) are considered in order to assign a score to each combination and then to choose the combination with the highest score. The main problem with this method is the combinatorial explosion it creates. Hence, the BF method is very difficult to apply in real conditions and moreover, makes the use of a longer analysis context impossible. To circumvent this problem, several approaches are possible. The first, called complete approaches, try to reduce the number of combinations by using pruning techniques and choice heuristics. In the context of WSD, a good example is the approach proposed by Hist and St-Onge (1998) that is based on lexical chains (a taxonomic semantic similarity measure based on the overall relations of WordNet) that combines restrictions during the construction of the global lexical chain with a greedy heuristic. According to Navigli (2009), the major problem of this approach is its lack of precision caused by the greedy strategy used. However, various improvements have been proposed among others by Silbert (2000). Other interesting complete approaches have been applied in the context of word sense disambiguation, in particular by Brody and Lapata (2008). The other approaches are called 'incomplete', as they explore only a part of the search-space using heuristics to guide them to areas that seem more promising. These heuristics are generally based on probabilities: choices are made stochastically.

Two main methods can be distinguished:

- neighborhood approaches (new configurations are created from existing configurations), among which are approaches from artificial intelligence such as genetic algorithms or optimization methods (e.g. simulated annealing);
- constructive approaches (new configurations are generated by iteratively adding solutions to the configurations under construction), among which are for example ant colony algorithms.

The reader may consult (Schwab et al. 2012) for more information.

3.3.2 Problem Configuration

To perform a Word Sense Disambiguation task, is to affect to each word w_i of a text of m words one of the senses of that word $w_{i,j}$. The definition of a sense j of word i is noted $d(w_{i,j})$. The search-space corresponds to all the possible sense combination for the text being processed. Therefore, a configuration C of the problem can be represented as an array of integers such that $j = C[i]$ is the selected sense j of w_i. For example, if we consider the simple text "*The mouse is eating cheese*", it has 3 words to be annotated ('*mouse*'; '*eat*'; '*cheese*'). If we consider the second sense for '*mouse*', the first sense for '*eat*'; and the third for '*cheese*', the configuration is [2;1;3].

4 New Ways to Evaluate Semantic Measures

While standard evaluation methods have been extremely useful for the development and improvement of the field of WSD, we have now reached a plateau in the development of such algorithms. Thus new ways of evaluating are required to go beyond that limit. We first set our working hypothesis and then go on and present the principles that governs our evaluation method.

4.1 Working Hypothesis

Two main working hypotheses have to be set in order to place an appropriate context for our work in this article.

4.1.1 Text as Context

In this article, we choose to consider a text in its entirety as the context window to be disambiguated. This choice is also made by Cowie (1992) for their WSD simulated annealing algorithm, made by Gelbukh et al. (2003) for their WSD genetic algorithm and the idea being taken up more recently by Navigli and Lapata (2010) and in our team Schwab et al. (2011, 2013, 2012). Many approaches, however, use a smaller context, especially for computational reasons, even if it is sometimes not explicitly reported. From our point of view, this leads to two problems. The first is that we have no way to ensure the consistency between the selected senses. Two generally incompatible senses can be chosen by the algorithm, because the context does not include the key word that can make the difference. For example, even with a window of six words before and six words after, the sentence "*The two planes were parallel to each other. The pilot had parked them meticulously*", "pilot" does not help to disambiguate the term "planes". The second problem is that texts usually hold some semantic unity. For example, as noted by Gale et al. (1992) or Hirst and St-Onge (1998), a word used several times in a text has generally the same sense; this information, better known as *one sense per discourse*, cannot be exploited within a windowed disambiguation context.

4.1.2 Uniform Global Score

The algorithms require some *fitness* measure to evaluate how good a configuration is. Even with the text as the context, it is possible to use several methods to compute a global score. For instance, one can weight relatively to the surrounding

words proportionally to their distance from that particular word. There are two approaches to a distance-based weighing:

- with respect to the distance in the number of interceding words
- with respect to the distance in a structure: syntactic structure, discourse structure, …

Such a criterion is important, however it is orthogonal to our object of study. We therefore chose a fixed weight of one for each word as a working hypothesis.

Hence, in this work, the score of the selected sense of a word is expressed as the sum of the local scores between that sense and the selected senses of all the other selected senses for the words of the text: for a full configuration, we simply sum the scores for all selected senses of the words of the text:

$$Score(C) = \sum_{i=1}^{m} \sum_{j=i}^{m} measure\left(w_{i,C[i]}, w_{j,C[j]}\right) \quad (1)$$

4.2 Principle

From a combinatorial optimization point of view, the ideal global score is a fitness value. In other terms, for a given local measure, the global score is an adequate estimator of the F1 score. This implies that the relationship between the global score and the F_1 should ideally be monotonic: the higher the global score, the higher the F1 measure. Various meta-heuristic approaches (simulated annealing, genetic algorithms,…) devise a heuristic global score function that exploits limited knowledge about the problem. The monotonicity prerequisite is often assumed to be true, as it is hoped that the global function will be a good estimator of the maximum a posteriori distribution of the optimal disambiguation. However, in truth, it is extremely difficult to construct a good estimator for the overall disambiguation of natural language texts. Such heuristics lead to biased and often noisy estimators, for which the monotonicity of the relationship between the global score and any form of score based on human judgement (for example F1 measure over a gold standard) can hardly be guaranteed.

Despite the very centrality of this issue, there has been little interest in the community to address such questions.

We study this problem through several measures that we can use to attempt to evaluate the adequacy of a given global score as a good estimator of the disambiguation of a text.

Starting from one or several texts extracted from a sense-annotated gold standard, the idea is to generate a sufficient[1] quantity of uniformly sampled configurations and to compute their F1 measure. Then, starting from this set, we can

[1] Sufficient in the sense of permitting the exhibition of statistical significance, even though in practice we generate several orders of magnitude more samples that the bare minimum necessary to obtain statistically significant differences in the average values.

compute, for each of its members the global score with one or more semantic measures. We can then represent the relationship between a configuration, its global score and F1 measure as a triple:

$$\langle C_i; F_i; S_i^{measure} \rangle \tag{2}$$

where C_i is the ith configuration of the dataset, F_i its corresponding F1 measure and $S_i^{measure}$ the corresponding global score for the measure *measure*.

From these triples, we can compute the measure we need to evaluate the appropriateness of the global score with relation to the F1 measure. We introduce here the notion of best attainable score and the correlation of the global score against the F1 measure.

4.2.1 Best Attainable Score

The best attainable score for a given global score formula, $Score_{best}^{measure}$, is the F1 measure that is obtained with the semantic measure *measure* so that the resulting global score is optimal. We assume the unicity of this value. Even though, in practice we should verify that it is indeed unique.

$$S_{best}^{measure} = \underset{F_i \in \langle C_i; F_i; S_i \rangle}{\operatorname{argmax}} \{score^{measure}(F_i)\} \tag{3}$$

4.2.2 Correlation Global Score/F1 Measure

Principle

As mentioned before, similarity-based WSD rest on the assumption that the global score is an adequate estimator of the F1 measure. The dynamics of such algorithms are based on maximizing the global score. We hence propose to evaluate a semantic measure through the correlation between the global score and the F1 measure. Of course the choice of the correlation measure defines different properties that we are trying to detect. For example a Pearson's correlation tests for can be used linear relationships, while the Spearman rank correlation coefficient imposes the weaker condition of monotonicity. In our case, we are just interested in ensuring monotonicity and thus, we use the Spearman rank correlation coefficient:

$$correlation(F, S) = r = \frac{\Sigma(F_i - \bar{F})(S_i - \bar{S})}{\sqrt{\Sigma(F_i - \bar{F})^2 \Sigma(S_i - \bar{S})^2}} \tag{4}$$

A correlation is a value between −1 and 1 with the following semantics:

- if the correlation is close to 1, the datasets are strongly correlated. In other word, there is a linear relationship between the distributions for the pearson correlation and an exact monotonic relationship in the case of the spearman coefficient. In simple terms when the value of A increases, the value of B as well.
- if the correlation is close to −1, the distributions are strongly inversely correlated and there exists an inverse monotonic or linear relationship between them. As A increases, B decreases, and vice versa.
- if correlation is around 0, there is no linear or monotonic relationship between the distributions.

An ideal measure would give a perfect monotonic relationship between the global score and the F1 measure. In other terms the correlation score would be 1.

How Representative is a Sample?

For our problem, a configuration is a vector of several hundred dimensions in the problem space. A score is assigned to each configuration. An optimal solution to the problem is a configuration with a score of 1. The search space is manifestly too large to explore exhaustively in search of an optimal solution. We, hence, sampled some configurations uniformly as an approximation of the whole search space. Of course we have no way of knowing if the sampling is representative, thus we adopted a technique that attempts to ensure that it is as representative as possible.

In simple terms, we divide the dataset in n different parts (with $n \geq 100$) and compute the correlation on each subset so as to be able to estimate if the variation in correlation is statistically significant over the total F1 measure. This exactly corresponds to a classical randomization test.

5 Construction of the Dataset

5.1 Gold Standard

Our dataset needs to be big enough to permit correlations that are statistically significant and well balanced between configurations with low F1 measure and configurations with high F1 measure to be computed. We choose to build the dataset from the first text of the Semeval 2007 task 7 coarse-grained all words corpus. This text is categorized by the task organiser as a news article, published in the Wall Street Journal.[2]

[2] The article is available here https://wiki.csc.calpoly.edu/CSC-581-S11-06/browser/trunk/treebank_paper/buraw/wsj_0105.ready.buraw.

In this text, there are 368 words to annotate. 66 of them are monosemic. Among the 302 remaining words, there is an average of 6.06 senses per words. We can then approximate the number of combinations as $6.06^{302} = 2 \times 10^{236}$.

The dataset will be accessible through the companion page of this article.[3]

5.2 Random Construction of Configurations

We randomly generated the configurations by starting from one of the configuration that obtained the best score (100 % of precision/recall/F1 measure). The idea of the generation algorithm is to randomly (uniformly) modify one or several senses in the configuration. Of course, it usually leads to new configurations with lower scores. Iterating this process several times permits to obtain configurations in the whole range of possible F1 measure values.

On this text, we obtained 1,910,332 configurations from which 1,184,125 were unique. Yet, the sample represents only 5.84×10^{-229} % of the search space. Figure 2 presents the likelihood density of our dataset, in function of the F1 measure. One can note that we don't have the same number of configurations for each possible F1 measure. It is not a problem, as it doesn't affect our method to have more configurations in one part of the F1 measure range, given that we cut our space in n different parts to compute our correlations. Moreover, it would be difficult and certainly impossible to obtain lot of configurations for the highest and the lowest F1 measure values.

6 First Experiments on Various Semantic Measures

6.1 Similarity Measures Evaluated

In this experiment we endeavour to evaluate and compare the global score function resulting from different similarity measures, and try to find commonalities and differences in terms of which one is the best estimator. In this particular experiment, we have our own implementation of the Lesk and Extended Lesk methods (respectively denoted GETALP-Lesk and GETALP-ExtLesk), and wish to compare them to the same measures as implemented in the `WN::Similarity` Perl package (denoted WNSIM-Lesk and WNSIM ExtLesk).[4] We will similarly consider its vector-based similarity measure (denoted WNSIM-Vectors).

[3] http://getalp.imag.fr/static/wsd/Schwab-et-al-SemanticSimilarity2014.html.

[4] http://wn-similarity.sourceforge.net.

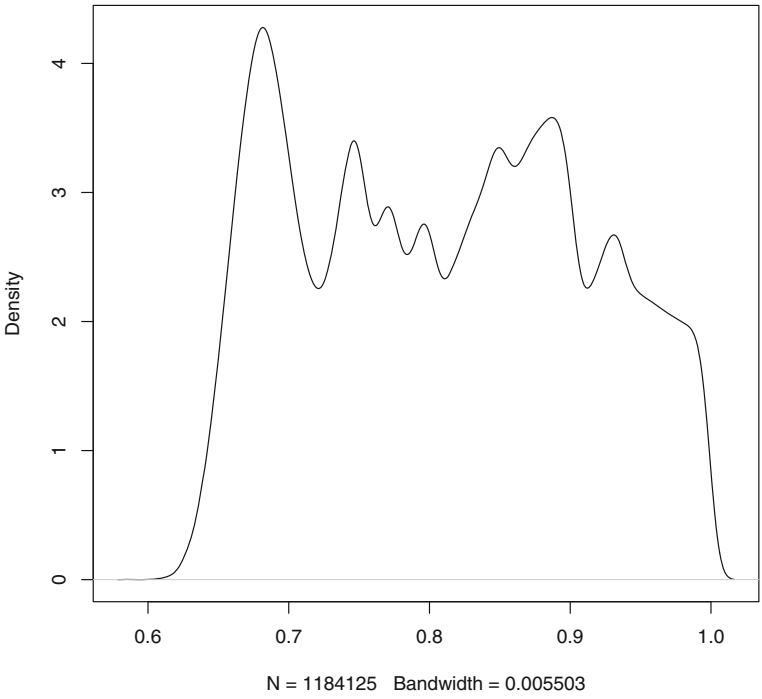

Fig. 2 Density of our dataset in function of the F1 measure

6.2 Best Atteignable Score

For each measure, we computed the best attainable scores on our dataset. Table 1 shows the corresponding F1 measure for each best global score.

First of all, we can note that Lesk obtains better results than Extended Lesk in both implementations. This is a quite surprising result. On previous known results with various gold standards and languages, Lesk always obtains worse results than Extended Lesks. It was the case with our implementation (Schwab et al. 2011), with Pedersen's implementation (Pedersen et al. 2005), with Baldwin et al.'s (2010) and with Miller et al.'s (2012).

It possibly means that global algorithms that use Lesk fall in a local maximum and are not able to find highest global scores. We will try to shed more light on this point.

If we compare the two implementations, Pedersen's appears to be better than our own, especially his WNSIM-Lesk, which obtains a very high F1 measure, 98.37 %, very close of 100 %. This result is certainly caused by squaring the overlap counted for the longest overlapping substring (see Sect. 6.1). We don't use that heuristic here, however, especially as it is computationally very expensive.

Table 1 Best atteignable scores of the measures on our dataset

Measure	Max score	Corresponding F1 measure
WNSIM-Lesk	360,793	0.9837
WNSIM-ExtLesk	2,804,835	0.8533
Getalp-Lesk	311,950	0.8533
Getalp-ExtLesk	3,555,160	0.8505
WNSIM-Vectors	43,795.3	0.8478

As shown in (Schwab et al. 2013), Pedersen's implementation's computational complexity is exactly $O(n \times m)$ while ours's complexity is, in the worst case, $O(n)$ with $n > m$.

6.3 Global Score/F1 Measure Correlation

In order to analyse and characterize the relationship between the global measure and the F1 measure, we now turn to the computation of a correlation measure between the global score and the F1 measure as explained in Sect. 4.2.2. Table 2 show this correlation calculated over the whole sampling of the search space. As we can see, the results vary wildly between the different measures, ranging from -0.2261 with our Extended Lesk, followed closely by Pedersen's Extended Lesk (-0.1755) and the vector similarity measure (-0.1664) all the way up to 0.9137 with Perdersen's Lesk measure. Our own Lesk measure has a correlation of 0.6968. Clearly, we see that some of the correlation are consistent with the F1 measure corresponding to maximal global scores, however, other correlation values are much more surprising. Indeed, if one considers the difference in correlation between ExtLesk and our own implementation of Lesk of $\delta_\rho = 0.4707$ and then the difference in the maximal F1 measure that is of only $\delta_F = 0.028$, the question of what explains such big correlation differences compared to the actual maximal F1 measure arises.

Table 2 Global score/F1 measure correlations on our dataset

Measure	Correlation global score/F1 measure
WNSIM-Lesk	0.9137
WNSIM-ExtLesk	−0.1755
Getalp-Lesk	0.6968
Getalp-ExtLesk	−0.2261
WNSIM-Vectors	−0.1664

Table 3 Correlations global score/F1 measure on our dataset

Measure	Min	1st quartile	Median	Mean	3rd quartile	Max
WNSIM-Lesk	0.9096	0.9124	0.9137	0.9136	0.9146	0.9170
Getalp-Lesk	0.6536	0.6924	0.6968	0.6960	0.7007	0.7135
WNSIM-Vectors	−0.1927	−0.1712	−0.1664	−0.1659	−0.1603	−0.1193
WNSIM-ExtLesk	−0.3898	−0.1815	−0.1755	−0.1776	−0.1697	−0.1462
Getalp-ExtLesk	−0.2493	−0.2326	−0.2261	−0.2270	−0.2233	−0.1370

Could this behaviour be the result of the convergence to a local maximum that is more difficult to escape with one measure over the other? Could it simply be that the correlation measure does not capture the relationship between the global score and the F1 measure? A large quantity of noise in the global score would certainly have a role to play in this discrepancy. If the monotonicity assumption between the scores is violated due to that potentially large presence of noise, can the correlation measures help identify the regions of the search space where the amount of noise is lesser and where the monotonicity assumption holds?

In order to have a better idea about the relationship between the global score and the F1 measure, it may be interesting to look at the distribution of correlation scores more closely, by first breaking down the distributions (Table 3).

Overall, the extrema of the distribution are relatively close to the mean. This seems to indicate a clearly significant difference[5] and thus does not give any further indications to explain what causes such a discrepancy between the maximal F1 measure and the correlation distribution.

Given that the correlation distribution encompasses the entire sampling, regardless of the F1 measure, it is difficult to draw any conclusion, thus we have broken down the correlation distribution depending on the F1 measure and represented it in a separate plot for each measure (Figs. 3, 4, 5, 6 and 7). Depending on the granularity of the increments in the F1 measure we would show more or less of the noise in the relationship. As we are interested in the general behaviour and not on minor artefacts due to noise, we selected a sliding window of 3 % F1 measure around each point of the plot. If the window size is any smaller, the noise makes any interpretation difficult and if the window size is any higher, interesting variations start being "smoothed away". The five lines on the plot, from top to bottom, respectively represent the maximum value, the 1st quartile, the mean, the 3rd quartile and the minimum value. Thus, we can have a more precise idea of the behaviours and of the distribution of correlation values for particular F1 measure neighbourhoods.

[5] Given that we have over a million configuration and that the correlation is calculated in chunks of 100 scores, each group contains over 10,000 samples, which at a 10^{-4} difference range should guarantee a sufficient statistical power.

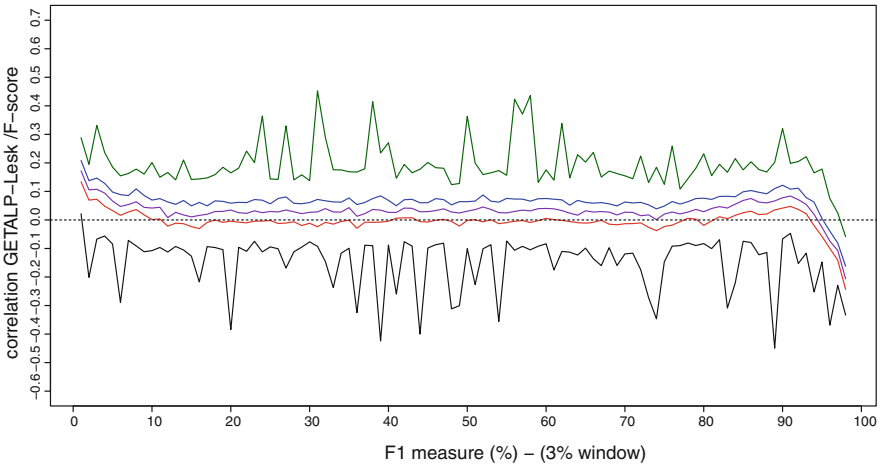

Fig. 3 Correlation of F1 measure and global score for our Lesk measure broken down by F1 measure in a 3 % sliding window

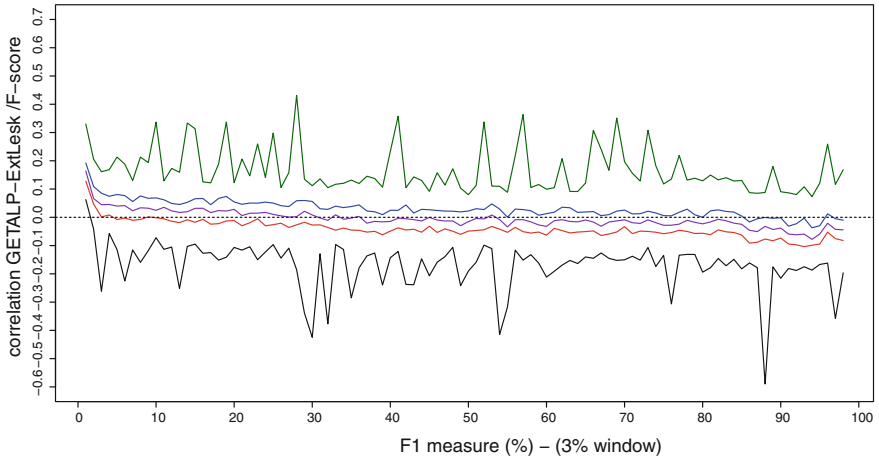

Fig. 4 Correlation of F1 measure and global score for our Extended Lesk measure broken down by F1 measure in a 3 % sliding window

We can immediately see that what explains the huge differences from earlier, is the position of the bulk of the distribution (inter-quartile portion) relative to the zero axis. Our Lesk implementation, for example exhibits consistently positive interquartile range throughout the F1 measure spectrum, while on the other hand, ExtLesk, WNSIM-ExtLesk or WNSIM-Vector the interquartile values are consistently negative.

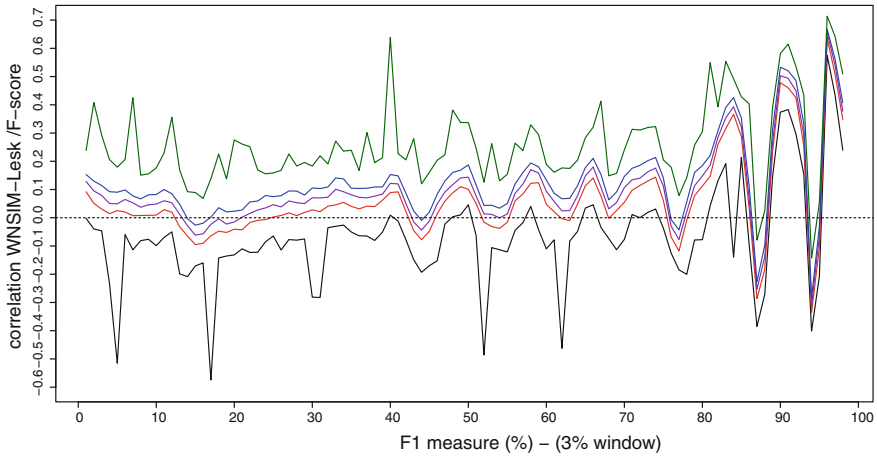

Fig. 5 Correlation of F1 measure and global score for the WNSim Lesk measure broken down by F1 measure in a 3 % sliding window

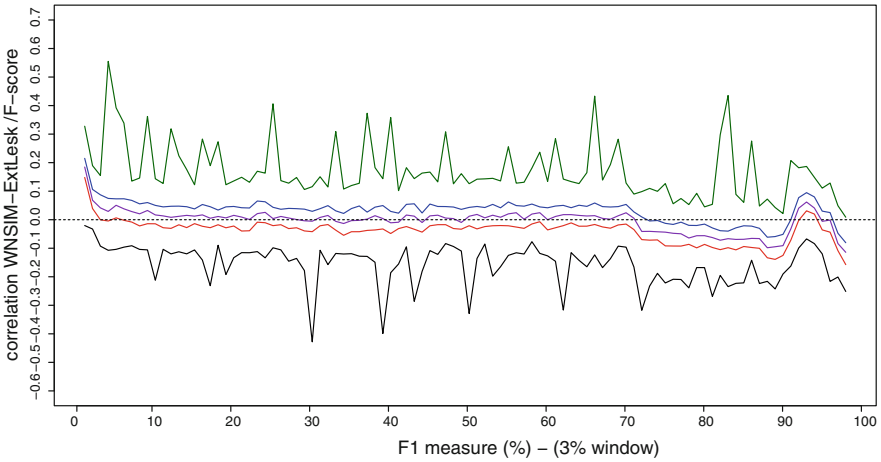

Fig. 6 Correlation of F1 measure and global score for the WNSim Extended Lesk measure broken down by F1 measure in a 3 % sliding window

One commonality between all the distributions is that there is a very wide divide between the maximal and minimal values that is roughly symmetric. In other words, when we have a correlation peak at the top of the distribution, there is often a matching low at the bottom of the distribution. Be it with GETALP-Lesk, GETALP-ExtLesk, WNSim-ExtLesk the maximum peaks tend to reach similar values in many places, this alone explain why we can get a good F1 measure despite a very negative overall correlation. The WNSim-Lesk and WNSim-ExtLesk consistently have a somewhat higher maximum peaks. Furthermore, the

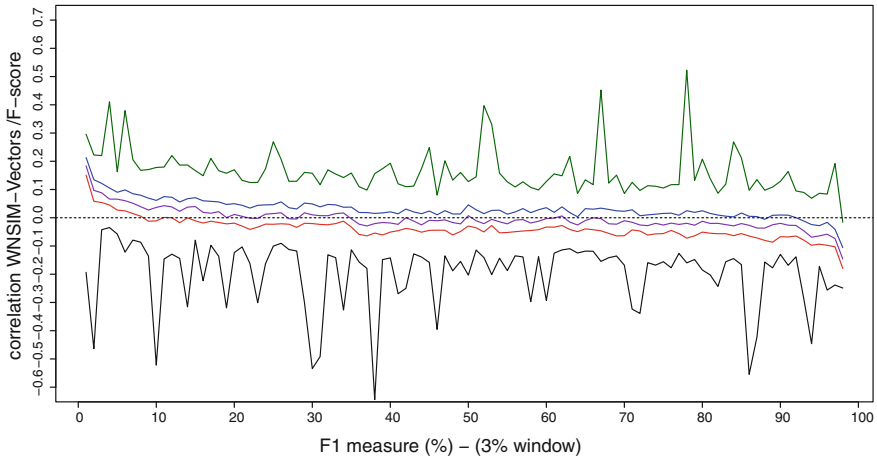

Fig. 7 Correlation of F1 measure and global score for the WNSim Vector similarity measure broken down by F1 measure in a 3 % sliding window

WNSIM-Lesk, features some very large variations towards the end of the F1 measure spectrum, but with a very narrow distribution of values (small inter-quartile range, small maximum/minimum spread), reaching as high as a correlation of 0.6 (around 0.80, 0.90, 0.95 F1 measure), but also very low in some places (around 0.85 and 0.93 F-measures). In contrast, however throughout the F1 measure spectrum, WNSim-Lesk has larger peaks both negative and positive and specifically, exhibits negative correlations in the lower end of the spectrum ($\sim$0.15) that may cause many more errors in WSD algorithms that have more chances on converging on a local minimum.

We can extend this observation overall to the distinction between Lesk and ExtLesk. The Lesk measures can reach potentially much better results that ExtLesk measures, however the search landscape is much more chaotic with Lesk and the noise certainly makes it much more challenging for any algorithm to find an optimal solution. On the other hand ExtLesk measures sacrifice potential maximally optimal configuration for a much smoother and consistent landscape, with much less minima or maxima traps. Thus the Lesk Extensions are merely a compromise to smooth the search space and make its exploration easier in exchange for lower potential scores. In practice, the trade-off is largely worth it as algorithms using extended Lesk yield much better results.

7 Conclusion

In our experiment, we sample the search space of WSD algorithms at the text level through different local similarity metrics summed into a global fitness score for a given configuration in the search space. Then, we move on to attempting to

characterize the properties and distribution of the global scores compared to a F1 measure computed with relation to a human annotated gold standard that constitutes a reference disambiguation. Since what interests us is the behaviour of the global function compared to the F1 measure and more specifically a monotonous relationship between the global score and the F1 measure, we compute the Spearman rank correlation coefficient, which quantifies exactly the relative behaviour of the two scores towards one another under an assumption of monotonicity. We then group the samples of our search space by 100 and compute a distribution of correlation values, which gives us an overall idea of the correlation. We then break down the correlation values with relation to the F1 measure in a 3 % sliding window, so as to perform a more fine-grained analysis. The analysis reveals that the main distinction between Lesk and Extended Lesk measures is that the Lesk measures have potentially much higher maximal scores and correlations, at the cost however, of much noisier search landscapes and numerous minima and maxima in the correlation values. Thus there are many variations and local non-monotonic behaviour that in turn makes the job of disambiguation algorithms much more challenging. In contrast, Extended Lesk measures incorporate information from related concepts from the taxonomy of a lexical resource and in a way potentially introduces more *linguistic noise*. On average, this leads to much lower correlations with the F1 measure, at the added benefit of greatly reducing the amount of variation in the correlation values and thus in the presence and density of local non-monotonous behaviour. As a result, WSD algorithms have a much easier time finding solutions and better than they would with the Lesk measure. Overall, the algorithms are very sensitive to local maxima, despite the numerous meta-heuristic countermeasures that they set in place and the landscape with a Lesk score is possibly among the most difficult search spaces for them.

We believe that our own findings give a promising window on what happens in WSD search spaces and open a new avenue of research towards their study and the improvement of search heuristics in the solving of WSD problems. We will continue to explore new ways of efficiently characterizing the search space and new improvements that leverage such findings and insight, in order to get even better disambiguation with knowledge-rich approaches.

References

Baldwin, T., Kim, S., Bond, F., Fujita, S., Martinez, D., & Tanaka, T. (2010). A reexamination of MRD-based word sense disambiguation. *ACM Transactions on Asian Language Information Processing*, *9*(1), 4:1–4:21. doi:10.1145/1731035.1731039, http://doi.acm.org/10.1145/1731035.1731039.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using wordnet. In *CICLing 2002*, Mexico City.

Brody, S., & Lapata, M. (2008). Good neighbors make good senses: Exploiting distributional similarity for unsupervised WSD. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*Coling 2008*), Manchester, UK (pp. 65–72).

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

Cowie, J., Guthrie, J., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *COLING 1992* (Vol. 1, pp. 359–365). Nantes, France.

Cramer, I., Wandmacher, T., & Waltinger, U. (2010). *WordNet: An electronic lexical database, chapter modeling, learning and processing of text technological data structures*. Heidelberg: Springer.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297–302.

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Fifth DARPA Speech and Natural Language Workshop* (pp. 233–237). Harriman, New York: États-Unis.

Gelbukh, A., Sidorov, G., & Han, S. Y. (2003). Evolutionary approach to natural language WSD through global coherence optimization. *WSEAS Transactions on Communications, 2*(1), 11–19.

Hirst, G., & St-Onge, D. D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.) *WordNet: An electronic lexical database* (pp. 305–332). Cambridge, MA: MIT Press.

Lesk, M. (1986). Automatic sense disambiguation using mrd: How to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC '86* (pp. 24–26). New York, NY, USA: ACM.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, *HLT '93* (pp. 303–308). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1075671.1075742, http://dx.doi.org/10.3115/1075671.1075742.

Miller, T., Biemann, C., Zesch, T., & Gurevych, I. (2012). Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of COLING 2012* (pp. 1781–1796). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from http://www.aclweb.org/anthology/C12-1109.

Navigli, R. (2009). WSD: A survey. *ACM Computing Surveys, 41*(2), 1–69.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science* (*SOFSEM*) (pp. 115–129).

Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 678–692.

Ng, H. T., & Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics, ACL '96* (pp. 40–47). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/981863.981869, http://dx.doi.org/10.3115/981863.981869.

Patwardhan, S., & Pedersen, T. (2006). Using wordnet based context vectors to estimate the semantic relatedness of concepts. In *EACL 2006 Workshop Making Sense of Sense—Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1–8).

Pedersen, T., Banerjee, S., & Patwardhan, S. (2005). Maximizing semantic relatedness to perform WSD. Research report, University of Minnesota Supercomputing Institute.

Pirró, G., & Euzenat, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, & B. Glimm (Eds.), *The semantic web—ISWC 2010* (Vol. 6496, pp. 615–630)., Lecture Notes in Computer Science Berlin/Heidelberg: Springer.

Rogers, D., & Tanimoto, T. (1960). A computer program for classifying plants. *Science, 132*(3434), 1115–1118.

Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97–123.

Schwab, D., Goulian, J., & Guillaume, N. (2011). Désambigusation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. In *Traitement Automatique des Langues Naturelles* (*TALN*), Montpellier, France.

Schwab, D., Goulian, J., & Tchechmedjiev, A. (2013). Worst-case complexity and empirical evaluation of artificial intelligence methods for unsupervised word sense disambiguation. *International Journal of Web Engineering* and *Technology 8*(2), 124–153. doi:10.1504/IJWET.2013.055713, http://dx.doi.org/10.1504/IJWET.2013.055713.

Schwab, D., Goulian, J., Tchechmedjiev, A., & Blanchon, H. (2012). Ant colony algorithm for the unsupervised word sense disambiguation of texts: Comparison and evaluation. In *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING 2012*), Mumbai (India).

Silber, H. G., McCoy, K. F. (2000). Efficient text summarization using lexical chains. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, *IUI '00* (pp. 252–255). New York, NY, USA: ACM.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352.

Wilks, Y., & Stevenson, M. (1998). Word sense disambiguation using optimised combinations of knowledge sources. In *COLING '98* (pp. 1398–1402). Stroudsburg, PA, USA: ACL. Retrieved from http://dx.doi.org/10.3115/980432.980797.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Zock, M., Ferret, O., & Schwab, D. (2010). Deliberate word access: An intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology*, *13*(4), 107–117. Retrieved from http://hal.archives-ouvertes.fr/hal-00953695.

Zock, M., & Schwab, D. (2011). Storage does not guarantee access: The problem of organizing and accessing words in a speaker's Lexicon. *Journal of Cognitive Science*, *12*, 233–258. Retrieved from http://hal.archives-ouvertes.fr/hal-00953672. (Impact-F 3.52 estim. in 2012).