CoTO: A Novel Approach for Fuzzy Aggregation of Semantic Similarity Measures

Jorge Martinez-Gil, Software Competence Center Hagenberg (Austria)

email: jorge.martinez-gil@scch.at, phone number: 43 7236 3343 838

Keywords: Knowledge-based analysis, Text mining, Semantic similarity measurement, Fuzzy logic

Abstract

Semantic similarity measurement aims to determine the likeness between two text expressions that use different lexicographies for representing the same real object or idea. There are a lot of semantic similarity measures for addressing this problem. However, the best results have been achieved when aggregating a number of simple similarity measures. This means that after the various similarity values have been calculated, the overall similarity for a pair of text expressions is computed using an aggregation function of these individual semantic similarity values. This aggregation is often computed by means of statistical functions. In this work, we present CoTO (Consensus or Trade-Off) a solution based on fuzzy logic that is able to outperform these traditional approaches.

1 Introduction

Textual semantic similarity measurement is a field of research whereby two terms or text expressions are assigned a score based on the likeness of their meaning [30]. Being able to accurately measure semantic similarity is considered of great relevance in many computer related fields since this notion fits well enough in a number of particular scenarios. The reason is that textual semantic similarity measures can be used for understanding beyond the literal lexical representation of words and phrases. For example, it is possible to automatically identify that specific terms (e.g., Finance) yields matches on similar terms (e.g., Economics, Economic Affairs, Financial Affairs, etc.) or an expert on the treatment of cancer could also be considered as an expert on oncology or tumor treatment.

The detection of different formulations of the same concept or text expression is a key method in a lot of computer-related disciplines. To name only a few, we can refer to a) data clustering where semantic similarity measures are necessary to detect and group the most similar subjects [4], b) data matching which consists of finding some data that refer to the same concept across different data sources [24], c) data mining where using appropriate semantic similarity measures can help to facilitate both the processes of text classification and pattern discovery in large texts [12], or d) automatic machine translation where the detection of terms pairs expressed in different languages but referring to the same idea is of vital importance [11].

Traditionally, this problem has been addressed from two different points of view: semantic similarity and relational similarity. However, there is a common agreement about the scope of each of them [3]. Semantic similarity states the taxonomic proximity between terms or text expressions [30]. For example, automobile and car are similar because they represent the same notion concerning means of transport. On the other hand, the more general notion of relational similarity considers relations between terms [31]. For example, nurse and hospital are related (since they belong to the healthcare domain) but they are far from represent the same real idea or concept. Due to its importance in many computer-related fields, we are going to focus on semantic similarity for the rest of this paper.

There are a lot of semantic similarity measures for identifying semantic similarity. However, the best results have been achieved when aggregating a number of simple similarity measures [13]. This means that after the various similarity values have been calculated, the overall similarity for a pair of text expressions is computed using an aggregation function of the individual semantic similarity values. This aggregation is often computed by means of statistical functions (arithmetic mean, quadratic mean, median, maximum, minimum, and so on) [22]. Our hypothesis is that these methods are not optimal, and therefore, can be improved. The reason is that these methods are following a kind of compensative approach, and therefore they are not able to deal with the non-stochastic uncertainty induced from subjectivity, vagueness and imprecision from the human language. However, dealing with subjectivity, vagueness and imprecision is exactly one of the major purposes of fuzzy logic. In this way, using techniques of this kind should help to outperform current results in the field of semantic similarity measurement. Therefore, the major contributions of this work can be summarized as follows:

- We propose CoTO (Consensus or Trade-Off), a novel technique for the aggregation of semantic similarity values that appropriately handles the non-stochastic uncertainty of human language by means of fuzzy logic.
- We evaluate the performance of this strategy using a number of general purpose and domain specific benchmark data sets, and show how this new approach outperforms the results from existing techniques.

The rest of this paper is organized as follows: Section 2 describes the state-of-the-art concerning semantic similarity measurement. Section 3 describes the novel approach for the fuzzy aggregation of simple semantic similarity measures. Section 4 describes our evaluations and the results that have been achieved. Finally, we draw conclusions and put forward future lines of research.

2 Related Work

The notion of textual semantic similarity represents a widely intuitive concept. Miller and Charles wrote: *...subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty* [26]. This viewpoint has been reinforced by other researchers in the field who observed that semantic similarity is treated as a property characterized by human perception and intuition [32]. In general, it is assumed that not only are the participants comfortable in their understanding of the concept, but also when they perform a judgment task they do it using the same procedure or at least have a common understanding of the attribute they are measuring [27].

In the past, there have been great efforts in finding new semantic similarity measures mainly due it is of fundamental importance in many application-oriented fields of the modern computer science. The reason is that these techniques can be used for going beyond the literal lexical match of words and text expressions. Past works in this field include the automatic processing of text and email messages [18], healthcare dialogue systems [5], natural language querying of databases [14], question answering [25], and sentence fusion [2].

On the other hand, according to Sanchez el al. [33]; most of these existing semantic similarity measures can be classified into one of these four main categories.

1. Edge-counting measures which are based on the computation of the number of taxonomical links separating two concepts represented in a given dictionary [19].

2. Feature-based measures which try to estimate the amount of common and non-common taxonomical information retrieved from dictionaries [29].

3. Information theoretic measures which try to determine similarity between concepts as a function of what both concepts have in common in a given ontology. These measures are typically computed from concept distribution in text corpora [17].

4. Distributional measures which use text corpora as source. They look for word co-occurrences in the Web or large document collections using search engines [6].

It is not possible to categorize our work into any of these categories. The reason is that we are not proposing a new semantic similarity measure, but a novel method to aggregate them so that individual measures can be outperformed. In this way, semantic similarity measures are like black boxes for us. However, there are several related works in the field of semantic similarity aggregation. For instance COMA, where a library of semantic similarity measures and friendly user interface to aggregate them are provided [13], or MaF, a matching framework that allow users to combine simple similarity measures to create more complex ones [21].

These approaches can be even improved by using weighted means where the weights are automatically computed by means of heuristic and meta-heuristic algorithms. In that case, most promising measures receive better weights. This means that all the efforts are focused on getting more complex weighted means that after some training are able to recognize the most important atomic measures for solving a given problem [23]. There are two major problems that make these approaches not very appropriate in real environments: First problem is that these techniques require a lot of training efforts. Secondly, these weights are obtained for a specific problem and it is not easy to find a way to transfer them to other problems. As we are going to see in the next section; CoTO, the novel strategy for fuzzy

aggregation of atomic measures that we present here, represents an improvement over traditional statistical approaches, and do not incur in the drawbacks from the heuristic and meta-heuristic ones, since it does not require any kind of training or knowledge transfer.

3 Fuzzy aggregation of semantic similarity measures

Currently, the baseline approach for computing the degree of semantic similarity between a pair of text expressions is based on an aggregation function of the individual semantic similarity values. This approach has proven to achieve very good results in practice. The idea is simple: to use quasi-linear means (like the median, the arithmetic mean, the geometric mean, the root-power mean, the harmonic mean, etc.) for getting the overall similarity score. In this way, we do not rely in an sole measure for taking important decisions. If there are some individual measures that do not perform very well for a given case, their effects are blurred by other measures that perform well. However, all these approaches present a major drawback: none of the operators is able to model in some understandable way an interaction between the different semantic similarity measures.

To overcome this limitation, first we develop a fuzzy membership function to capture the importance of different semantic similarity measures, and then use an operator for aggregation of multiple similarity measures corresponding to different features of semantic similarity. Experimental evaluations included in the next section will confirm the suitability of the proposed method.

3.1 Fuzzy modeling of semantic similarity

During a long time, similarity in general and semantic similarity in particular have been unknown and intangible attributes for the research community. According to O'Shea et al. the question that had to be faced was: Is similarity just some vague qualitative concept with no real scientific significance? [27]. To answer the question a broad survey of the literature, taking in as many fields as possible, was conducted. This revealed a generalized abstract theory of similarity [34], tying in with well-respected principles of measurement theory, many uses as both a dependent and independent variable in the fields of Cognitive Science, Neuropsychology and Neuroscience, and many practical applications.

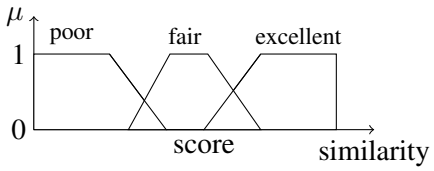

Figure 1: Fuzzy degrees of semantic similarity using three linguistic terms. Please note that, in this case, each linguistic value can belong (to some extent) to two different linguistic terms

Traditionally, a semantic similarity measure is defined as a function μ_1 x $\mu_2 \rightarrow$ R that associates the degree of correspondence for the entities μ_1 and μ_2 to a score $s \in$ R in the range $[0, 1]$, where a score of 0 states for not correspondence at all, and 1 for total correspondence of the entities μ_1 and μ_2. However, in fuzzy logic, linguistic values and expressions are used to describe numbers used in conventional systems. For example, the terms "low" or "wide-open" are designated as linguistic terms of the values "temperature" or "heating valve opening". If an input variable is described by linguistic terms, it is referred to as a linguistic value.

Each linguistic term is described by a Fuzzy Set M. It is defined mathematically by the two statements basic set G and membership function μ. The membership function states the membership of every element of the universe of discourse G (e.g. numerical values of a time scale [age in years]) in the set M (e.g. old) in the form of a numerical value between zero and one. If the membership function for a specific value is one, then the linguistic statement corresponding to the linguistic term applies in all respects (e.g. old for an age of 80 years). If, in contrast, it is zero, then there is absolutely no agreement (e.g. "very young" for an age of 80 years).

Since most fuzzy sets have a universe of discourse consisting of the real line R, it would be impractical to list all the pair defining a membership function. A more convenient and concise way to define a membership function is to express it as a mathematical formula. This can be expressed by means of the following equation. The parameters a, b, c, d (with $a < b <= c < d$) determine the x coordinates of the four boundaries of the underlying membership function.

$$m(x; a, b, c, d) = max\left(min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right)$$

In our case, we have three linguistic terms for assessing the degree of semantic similarity between two terms or text expressions: bad, fair and good[1]. Our membership function states the membership of each of these linguistic terms in the form of a trapezoid bounded between zero and one. Figure 1 shows us this more clearly: each linguistic value can belong to one of the three linguistic terms. Sometimes, a given linguistic value can belong (to some extent) to two or more different linguistic terms. For example, the semantic similarity for the word pair vehicle-motorbike can be assessed as 0.4 fair and 0.6 good (maybe 4 experts said fair and 6 experts said good). This fact allows us to model semantic similarity in a non-compensative way, thus, a much more flexible way that traditional approaches. As a result, more sophisticated aggregation schemes can be proposed.

3.2 Fuzzy aggregation of atomic measures

In the field of semantic similarity measurement, aggregation functions are generally defined and used to combine several numerical values (from the different semantic similarity measures to be aggregated) into a single one, so that the final result of the aggregation takes into account all the individual values in a given manner. The fundamental similarity measures which cover many specific characteristics from text strings are the most widely used measures in state of the art. However, the real issue arises when these similarity measures give different results for the same scenario. Different techniques have been used to aggregate the results of different similarity measures. Most of them have reached a high level of success [22].

In fuzzy logic, things are a little bit different. Values can belong to either single numerical or non numerical scale, but the existence of a weak order relation on the set of all possible values is the minimal requirement which has to be satisfied in order to perform aggregation. Nevertheless, the values to be aggregated belong to numerical scales, which can be of ordinal or cardinal type. Once values are defined, it is possible to aggregate them and obtain new value defined on the same scale, but this can be done in many different ways according to what is expected from the aggregation operation, what is the nature of the values to be aggregated, and what kind of scale has been used [15].

[1]We will investigate approaches using a larger amount of linguistic terms in the future

It is necessary to remark that aggregation is a very extensive research field in which numerous types of aggregation functions or operators exist. They are all characterized by certain mathematical properties and aggregate in a different manner. But in general, aggregation operators can be divided into three categories [16]: conjunctive, disjunctive and compensative operators.

- Conjunctive operators combine values as if they were related by a logical AND operator. That is, the result of combination can be high only if all the values are high.
- Disjunctive operators combine values as an OR operator, so that the result of combination is high if at least one value is high.
- Compensative operators which are located between min and max (bounds of the t-norm and t-conorm families). In this kind of operators, a bad (good) score on one criterion can be compensated by a good (bad) one on another criterion, so that the result will be medium.

After previous research in the field of statistical aggregation of semantic similarity measures, we realize that existing approaches are always based on compensative operators. However, in this work we decided to investigate what happens if dissident values are not taken into account for computing the overall score. The rational behind this idea is that if dissident values are not good, taking into account them may decrease the quality the overall similarity score. On the contrary, if dissident values are correct, ignoring them can be detrimental. Our intuition is that consensus will be right most of times (atomic semantic similarity measures to be aggregated are supposed to be good), and therefore this strategy should produce more good than harm, but only a rigorous evaluation using well-known benchmark data sets could verify this.

Therefore, our proposal is based on the idea of **Consensus or Trade-off** what means that atomic semantic similarity measures have to be aggregated without reflecting dissident recommendations in case of a consensus have been reached or using a high degree of trade-off in case a recommendation consensus from atomic measures does not exist. The problem in applying this is that an appropriate fuzzy aggregation operator for implementing this strategy does not exist. For this reason, we have to design it by means of IF-THEN rules.

To be more formal, our CoTo aggregation operator on a fuzzy set $(2 \geq n)$ is defined by a function

$$h : [0,1]^n \rightarrow [0,1]$$

which follows these axioms:

- *Boundary condition:* $h(0,0,...,0) = 0$ and $h(1,1,...,1) = 1$
- *Monotonicity:* For any pair $\langle a_1, a_2, ..., a_n \rangle$ and $\langle b_1, b_2, ..., b_n \rangle$ of n-tuples such that $a_i, b_i \in [0,1]$ for all $i \in N_n$, if $a_i \leq b_i$ for all $i \in N_n$, then $h(a_1, a_2, ..., a_n) \leq h(b_1, b_2, ..., b_n)$; that is, h is monotonic increasing in all its arguments.
- *Continuity*: *h* is a continuous function.

We need a fuzzy associative matrix for implementing our strategy. A fuzzy associative matrix expresses fuzzy logic rules in tabular form. These rules take n variables as input, mapping cleanly to a vector. Linguistic terms are bad (the two text entities to be compared are not similar at all), fair (the two text entities to be compared are moderately similar) and excellent (the two text entities to be compared are very similar). A linguistic term reaches a consensus when it receives the highest number of votes, in that case its associated fuzzy set will be the result of the aggregation process. In case, two or more linguistic terms may receive the same major amount of votes[2], two or more fuzzy sets will be combined in a desirable way to produce a single fuzzy set. This is exactly the purpose of our CoTo aggregation operation. Our final overall score will be computed by means of the trade-off of their respective associated fuzzy sets. This trade-off can be achieved by any of the traditional processes of producing a quantifiable result by means of defuzzification.

Even once the fuzzy model has been defined, it is necessary to configure some parameters concerning the fuzzy terms from the model. This means that it is necessary to perform a parametric study about the degree of overlapping between trapezoids, number of linguistic terms, defuzzification method, etc. for deciding when a pair of text expressions is going to be considered or not semantically equivalent.

[2]For example, a scheme with 5 semantic similarity measures, where bad receives 1 vote, fair receives 2 votes and excellent receives 2 votes

This can be achieved by means of parameter tuning and refinement. Parameter tuning consists of optimizing the internal running parameters in order to maximize the fulfillment of our goal (replicate the human behavior when assessing the semantic similarity of each of the expression pairs contained in the benchmark data sets we are going to solve). In this case, we refer to the the maximization of efficiency (or error minimization) so that the benchmark data set can be solved with a minimum number of errors. This fact is of vital importance since the smaller number of errors, the greater the quality of the results obtained by our strategy.

For the defuzzification process, we have chosen the method Center of Gravity (CoG) (a.k.a. fuzzy centroid method) to find the final non-fuzzy value associated with the semantic similarity between the text expressions to be compared. This classical method consists of computing the center of gravity for the area under the curve determined by the rules triggered, and we have chosen it because this method represents a trade-off between the rules triggered (what it is exactly the Trade-Off we have mentioned along the manuscript). This method can be computed as it is expressed in the following formula:

$$CoG = \frac{\sum_{x=a}^{b} \mu_A(\chi)x}{\sum_{x=a}^{b} \mu_A(\chi)}$$

This method is similar to the formula for calculating the center of gravity in physics. The weighted average of the membership function or the center of the gravity of the area bounded by the membership function curve is computed to be the most crisp value of the fuzzy quantity.

Figure 2 shows us a summary of the whole process for clarification purposes. This process starts by encoding a numerical value into a linguistic term by matching the given value within the limits of the existing fuzzy sets these linguistic terms represent. Then, each linguistic term serve as an input for the rule engine which implements the aggregation operator (CoTO). One of the advantages of fuzzy logics is that the design of complex rules engines become an intuitive task (mainly due to the proximity of the linguistic terms to natural language). In a further step, the rule engine triggers the rules that configure the resulting fuzzy set. Finally, the final aggregated score is retrieved by computing the CoG of the resulting fuzzy set.

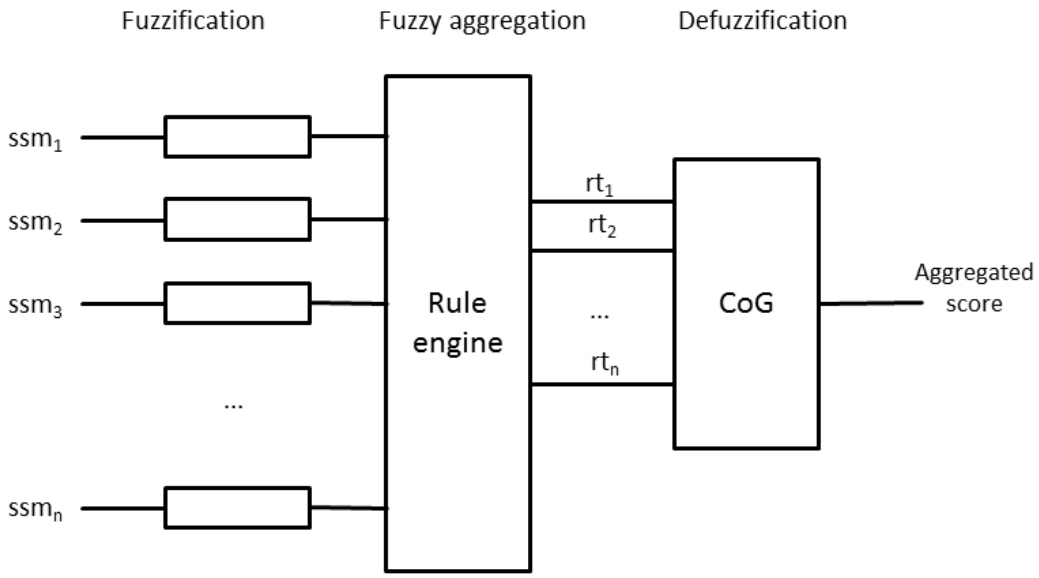

Figure 2: Overall summary of the fuzzy aggregation process: a) Values from n semantic similarity measures (ssm) are fuzzificated into linguistic terms, b.1) a rule engine determines if there is a consensus between linguistic terms and triggers n rules (rt) accordingly, b.2) if there is a consensus, only a resulting fuzzy set will be generated, if not, two o more fuzzy sets representing the (two or more) most voted choices will be generated, c) the CoG of the resulting set(s) is computed

4 Evaluation

The comparison between the aggregated value for semantic similarity measures and human similarity judgments is going to be calculated in terms of correlation between the two sets of ratings, thereby giving us a qualitative assessment of the correlation with human similarity judgments. This in turn is an indication of the usefulness in, for example, an information retrieval task. In this section we summarize the main experiments and the results obtained in our study. We have used three different benchmark data sets. Firstly, we aim to measure semantic similarity for general terms, to do that we are going to use the Miller-Charles benchmark data set [26] which is intended for measuring the quality of artificial techniques when assessing the semantic similarity of general words. Secondly, we are going to test our approach using two domain specific benchmark data sets from the biomedical field. First of them is called the Biomedical Medical Subject Headings (MeSH) [28] and it is intended for measuring the quality of artificial techniques when assessing the semantic similarity of very specific words belonging to the field of the biomedicine. The second one is a benchmark data set concerning medical disorders that was created by Pedersen et al. in collaboration with Mayo Clinic experts [28]. Finally, we discuss the result from our experiments.

It is important to remark, that our technique is going to be compared to baseline aggregation methods. This baseline strategy consists of using the following family of means:

$\bar{x}(m) = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^m\right)^{\frac{1}{m}}$

By choosing different values for the parameter m, the following types of means are obtained: $m \to \infty$ maximum, $m = 2$ quadratic mean, $m = 1$ arithmetic mean, $m \to 0$ geometric mean, $m = -1$ harmonic mean, $m \to -\infty$ minimum. It is also necessary to explain that all tables, except those for the Miller & Charles ratings, are normalized into values in [0, 1] range for ease of comparison. This means that we cannot include geometric and harmonic means since we allow the value 0 when assessing semantic similarity and this may involve a error concerning division by zero.

In summary, from a strictly mathematical point of view, solving this problem consists on obtaining the maximum value for the Pearson Correlation Coefficient [1] of two numeric vectors, one generated by human experts and other generated by a computational algorithm. The final result can vary between

-1 (results from humans and the proposed algorithm are exactly the opposite) to 1 (results from humans and the proposed algorithm are exactly the same). Obviously, our challenge is to obtain a score of 1 what may mean that our solution is able to perfectly replicate human behavior. It is also important to remark that we compute the p-value for each result. The p-value is the value representing the probability to find the given result if the correlation coefficient were in fact zero (null hypothesis). If this probability is lower than the conventional $5.0 \cdot 10^{-2}$, then the correlation coefficient can be considered as statistically significant.

4.1 General purpose data set

First experiment is performed by using the Miller-Charles benchmark data set [26] which is a widely used reference data set for evaluating the quality of new semantic similarity measures for word pairs. The rationale behind this way to evaluate quality is that each result obtained by means of artificial techniques may be compared to human judgments. Therefore, the ultimate goal is to replicate human behavior when solving tasks related to semantic similarity without any kind of supervision. Table 1 shows us the 30 word pairs of the data set. The columns called *WordA* and *WordB* represent the word pairs belonging to the Miller-Charles benchmark data set. This collection of word pairs ranges from words which are not similar (for instance, rooster-voyage) to word pairs that are synonyms according to human judgment (for instance, automobile-car). Column called *Human* represent the opinion provided by people. This opinion was originally given in numeric score in the range [0, 4] where 0 stands for no similarity between the two words from the word pair and 4 stands for complete similarity. There is no problem when assessing semantic similarity using values belonging to the interval [0, 1] since the Pearson correlation coefficient is invariant against a linear transformation.

For the first experiment, we aim to smartly aggregate semantic similarity measures between terms which consists of using dictionaries. These measures are: a) Hirst, b) Jiang, c) Resnik, d) Leacock and e) Lin. A detailed description of these measures is out the scope of this work, but some explanatory insights are described in [8]. For us, it is enough to know these single measures are the state-of-the-art in the field of semantic similarity measurement [9].

WordA	WordB	Human
rooster	voyage	0.08
noon	string	0.08
glass	magician	0.11
cord	smile	0.13
coast	forest	0.42
lad	wizard	0.42
monk	slave	0.55
forest	graveyard	0.84
coast	hill	0.87
food	rooster	0.89
monk	oracle	1.10
car	journey	1.16
brother	lad	1.66
crane	implement	1.68
brother	monk	2.82
implement	tool	2.95
bird	crane	2.97
bird	cock	3.05
food	fruit	3.08
furnace	stove	3.11
midday	noon	3.42
magician	wizard	3.50
asylum	madhouse	3.61
coast	shore	3.70
boy	lad	3.76
journey	voyage	3.84
gem	jewel	3.84
automobile	car	3.92

Table 1: Miller-Charles benchmark data set. Human ratings are between 0 (not similar at all) and 4 (totally similar)

Method	**Score**	**p-value**
Hirst	0.69	$2.5 \cdot 10^{-5}$
Jiang	0.70	$1.7 \cdot 10^{-5}$
Resnik	0.77	$3.3 \cdot 10^{-7}$
Leacock	0.82	$1.0 \cdot 10^{-8}$
Lin	0.82	$1.0 \cdot 10^{-8}$
Minimum	0.66	$3.6 \cdot 10^{-5}$
Midrange	0.78	$1.9 \cdot 10^{-7}$
Median	0.80	$6.0 \cdot 10^{-8}$
Quadratic mean	0.81	$3.0 \cdot 10^{-8}$
Arithmetic mean	0.81	$3.0 \cdot 10^{-8}$
Maximum	0.82	$1.0 \cdot 10^{-8}$
CoTO	**0.85**	$4.0 \cdot 10^{-9}$

Table 2: Results for the aggregation of the different semantic similarity measures based on measures taking advantage of dictionaries. Some baseline strategies are able to reduce the risk of using only one measure for taking decisions. However, CoTO beats all simple similarity measures and compensative operators. Moreover, the results are statistically significant.

Table 2 shows the results for the aggregation of the different semantic similarity measures based on dictionary measures. Some traditional aggregation strategies (baseline approach) are in line with the best single measures, but the best score is achieved by using CoTO. It is also important to remark that the results we have achieved are statistically significant.

In our second experiment, we aim to determine the degree of semantic similarity between terms which consists of using the knowledge inherent in the historical search logs from the Google search engine. We have decided to perform our first experiment using four semantic similarity measures exploiting historical search patterns on the Web [20]. These semantic similarity measures are: a) frequent co-occurrence of terms in search patterns, b) computation of the relationship between search patterns, c) outlier coincidence in search patterns, and d) forecasting comparisons. Each of these semantic similarity

Method	**Score**	**p-value**
Pearson	0.13	$5.1 \cdot 10^{-1}$
Forecast	0.19	$3.3 \cdot 10^{-1}$
Co-occur.	0.36	$6.0 \cdot 10^{-2}$
Outlier	0.37	$5.2 \cdot 10^{-2}$
Maximum	0.23	$2.4 \cdot 10^{-1}$
Midrange	0.31	$1.1 \cdot 10^{-1}$
Minimum	0.32	$1.0 \cdot 10^{-1}$
Quadratic mean	0.38	$4.6 \cdot 10^{-2}$
Arithmetic mean	0.44	$1.9 \cdot 10^{-2}$
Median	0.46	$1.4 \cdot 10^{-2}$
CoTO	**0.52**	$4.6 \cdot 10^{-3}$

Table 3: Results for the aggregation of the different semantic similarity measures based on measures taking advantage of Google historical data. Some baseline aggregation strategies outperform single measures. However, CoTO beats all simple similarity measures and compensative operators. Moreover, the results are statistically significant.

measures is a distributional measure. The reason is these measures try to determine the likeness between terms by means of a smart analysis of their occurrences in the historical web search logs from Google.

Table 3 shows the results for the aggregation of the different semantic similarity measures based on measures taking advantage of Google historical data. Some traditional aggregation strategies (Quadratic mean, Arithmetic mean and Median) outperform single measures, but the best score is achieved, once again, by using CoTO.

Now we propose a new experiment using another kind of semantic similarity measure: the Normalized Google Distance [10]. This semantic similarity measure consists of computing the the number of hits returned by the Google search engine for a given set of keywords. The rationale behind this is that terms with similar meanings in a natural language sense tend to be close in units of Normalized Google Distance, while words with dissimilar meanings tend to be farther apart. This approach only uses the

Method	**Score**	**p-value**
Ask	0.26	$1.8 \cdot 10^{-1}$
Yahoo!	0.34	$7.7 \cdot 10^{-2}$
Bing	0.43	$2.2 \cdot 10^{-2}$
Google	0.47	$1.1 \cdot 10^{-2}$
Maximum	0.26	$1.8 \cdot 10^{-1}$
Midrange	0.32	$9.9 \cdot 10^{-2}$
Minimum	0.42	$2.6 \cdot 10^{-2}$
Quadratic mean	0.53	$3.7 \cdot 10^{-3}$
Arithmetic mean	0.61	$5.7 \cdot 10^{-4}$
Median	0.61	$5.7 \cdot 10^{-4}$
CoTO	**0.64**	$2.4 \cdot 10^{-4}$

Table 4: Results for the aggregation of the different semantic similarity measures based on Google Distance over web search engines. Some baseline aggregation strategies outperform single measures. But CoTO beats all simple measures and compensative operators.

probabilities of search terms extracted from the web corpus in question. Assuming that *x* and *y* are the terms to be compared, the formula is:

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

Additionally, to perform this experiment we are using also other web search engines Ask, Bing and Yahoo!. This idea was introduced in [22]. Then, we are going to aggregate the values using our family of means and to compare the results with our CoTO strategy.

Table 4 shows the results for the aggregation of the different s measures based on Google Distance over popular web search engines (Ask, Yahoo!, Bing and Google). Once again, some baseline aggregation strategies (Quadratic mean, Arithmetic mean and Median) outperform single measures. But once again, CoTO beats all simple semantic similarity measures and compensative operators.

4.2 Domain specific data sets

MeSH, the first of the biomedical benchmark data sets is composed by a set of 36 word pairs extracted from the MeSH data set [28]. Table 5 shows us a part of this data set. The columns called *ExpressionA* and *ExpressionB* represent the text expressions belonging to this benchmark data set. Column called *Human* represent the opinion provided by 8 medical experts. The similarity between text expressions have been also assessed between 0 and 1. Therefore, this data set ranges from biomedical expressions which are not similar (for instance, Anemia-Appendicitis) to expression pairs that are synonyms according to expert judgment (for instance, Antibiotics-Antibacterial Agents).

Table 6 shows the results for the aggregation of the different semantic similarity measures based on cutting-edge similarity measures from the biomedical domain. Explaining each of them is out of the scope of this work, but a detailed description can be found in [9]. Once again, the strategy CoTo (Consensus or Trade-Off) is able to beat all the single measures as well as all the compensative operators by a wide margin.

Concerning the second benchmark data set from the biomedical domain. It was created by Pedersen et al. in collaboration with experts from the Mayo Clinic [28]. Table 7 shows us this data set. We have to say that the columns called *ExpressionA* and *ExpressionB* represent the expressions pairs belonging to the Pedersen-Mayo Clinic benchmark data set. This collection of 30 text expressions ranges from cases which are not similar (for instance, Hyperlidpidemia-Metastasis) to other cases that are synonyms according to the expert judgment (for instance, Renal failure-Kidney failure). Column called *Human* represents the rating provided by experts from the biomedical domain.

Table 8 shows the results for the aggregation of the different semantic similarity measures based on cutting-edge similarity measures. Detailed description for each of these algorithms can be found in [9]. Once again, we have that the strategy CoTO (Consensus or Trade-Off) is able to beat, once again, all the single measures as well as all the compensative operators by a wide margin.

ExpressionA	ExpressionB	Human
Anemia	Appendicitis	0.031
Otitis Media	Infantile Colic	0.156
Dementia	Atopic Dermatitis	0.060
Bacterial Pneumonia	Malaria	0.156
Osteoporosis	Patent Ductus Arteriosus	0.156
Sequence	Antibacterial Agents	0.155
A. Immunno. Syndrome	Congenital Heart Defects	0.060
Meningitis	Tricuspid Atresia	0.031
Sinusitis	Mental Retardation	0.031
Hypertension	Failure	0.500
Hyperlipidemia	Hyperkalemia	0.156
Hypothyroidism	Hyperthyroidism	0.406
Sarcoidosis	Tuberculosis	0.406
Psychology	Cognitive Science	0.593
Anemia	Deficiency Anemia	0.437
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial Ischemia	Myocardial Infarction	0.750
Hepatitis B	Hepatitis C	0.562
Carcinoma	Neoplasm	0.750
Pulmonary Stenosis	Aortic Stenosis	0.531
Failure to Thrive	Malnutrition	0.625
Breast Feeding	Lactation	0.843
Antibiotics	Antibacterial Agents	0.937
Seizures	Convulsions	0.843
Pain	Ache	0.875
Malnutrition	Nutritional Deficiency	0.875
Measles	Rubeola	0.906
Chicken Pox	Varicella	0.968
Down Syndrome	Trisomy 21	0.875

Table 5: MeSH biomedical benchmark. Human ratings are between 0 (not similar at all) and 1 (totally similar)

Method	**Score**	**p-value**
Li	0.707	$7.2 \cdot 10^{-7}$
J&C	0.718	$4.1 \cdot 10^{-7}$
Lin	0.718	$4.1 \cdot 10^{-7}$
Resnik	0.721	$3.5 \cdot 10^{-7}$
Maximum	0.711	$5.9 \cdot 10^{-7}$
Minimum	0.712	$5.6 \cdot 10^{-7}$
Median	0.716	$4.6 \cdot 10^{-7}$
Arithmetic mean	0.722	$3.3 \cdot 10^{-7}$
Midrange	0.724	$3.0 \cdot 10^{-7}$
Quadratic mean	0.725	$2.0 \cdot 10^{-8}$
CoTO	**0.771**	$7.1 \cdot 10^{-9}$

Table 6: Results for the aggregation of the different semantic similarity measures based on cutting-edge similarity measures. The strategy CoTO (Consensus or Trade-Off) is able to beat all the single measures as well as all the compensative operators by a wide margin. The results are statistically significant. Moreover, the results are statistically significant.

ExpressionA	ExpressionB	Human
Renal failure	Kidney failure	1.000
Heart	Myocardium	0.750
Stroke	Infarct	0.700
Abortion	Miscarriage	0.825
Delusion	Schizophrenia	0.550
Congestive heart failure	Pulmonary edema	0.350
Metastasis	Adenocarcinoma	0.450
Calcification	Stenosis	0.500
Diarrhea	Stomach cramps	0.325
Mitral stenosis	Atrial fibrillation	0.325
C. pulmonary disease	Lung infiltrates	0.475
Rheumatoid arthritis	Lupus	0.275
Brain tumor	Intracranial hemorrhage	0.325
Carpel tunnel syndrome	Osteoarthritis	0.275
Diabetes mellitus	Hypertension	0.250
Acne	Syringe	0.250
Antibiotic	Allergy	0.300
Cortisone	Total knee replacement	0.250
Pulmonary embolus	Myocardial infarction	0.300
Pulmonary fibrosis	Lung cancer	0.350
Cholangiocarcinoma	Colonoscopy	0.250
Lymphoid hyperplasia	Laryngeal cancer	0.250
Multiple sclerosis	Psychosis	0.250
Appendicitis	Osteoporosis	0.250
Rectal polyp	Aorta	0.250
Xerostomia	Alcoholic cirrhosis	0.250
Peptic ulcer disease	Myopia	0.250
Depression	Cellulites	0.250
Varicose vein	Entire knee meniscus	0.250
Hyperlidpidemia	Metastasis	0.250

Table 7: Mayo Clinic biomedical benchmark. Human ratings are between 0 (not similar at all) and 1 (totally similar)

Method	**Score**	**p-value**
Jcn	0.111	$2.8 \cdot 10^{-1}$
Wup	0.483	$3.4 \cdot 10^{-3}$
Hso	0.701	$8.0 \cdot 10^{-6}$
Path	0.753	$7.9 \cdot 10^{-7}$
Minimum	0.354	$2.7 \cdot 10^{-2}$
Maximum	0.483	$3.4 \cdot 10^{-3}$
Midrange	0.501	$2.4 \cdot 10^{-3}$
Quadratic mean	0.667	$2.8 \cdot 10^{-5}$
Arithmetic mean	0.747	$1.0 \cdot 10^{-6}$
Median	0.786	$1.3 \cdot 10^{-7}$
CoTO	**0.799**	$6.0 \cdot 10^{-8}$

Table 8: Results for the aggregation of the different semantic similarity measures based on cutting-edge similarity measures. The strategy CoTO (Consensus or Trade-Off) is able to beat all the single measures as well as all the compensative operators by a wide margin. Moreover, the results are statistically significant.

4.3 Discussion

Results show us that our CoTO strategy is able to consistently beat existing approaches based on compensative operators when solving both general purpose and domain specific data sets. In fact, CoTO has outperformed all existing semantic similarity measures and aggregation methods in all experiments performed in this study. Moreover, the results obtained were statistically significant. The reason is that unlike baseline aggregation techniques based on compensative operators, this aggregation strategy requires a consensus or at least a trade-off between between majority opinions. This means that dissident votes are not taken into account to compute the overall semantic similarity score. Therefore, our initial hypothesis seems to be true: dissident values may decrease the final quality of the overall semantic similarity score, so this fact can help to outperform current aggregation techniques based on compensative operators.

The major reason for getting these good results is that fuzzy logic and the classic compensative approach try to address different forms of uncertainty. Whereas both fuzzy logic and the classic compensative approach can represent degrees of certain kinds of subjective judgment, CoTO uses the concept of fuzzy set membership, i.e., how much a variable is in a set (there is not necessarily any uncertainty about this degree), and the classic compensative approach uses the concept of subjective probability, i.e., how probable is it that a variable is in a set. The technical consequence of this distinction is that CoTO relaxes the axioms of the classical compensative approach, which are derived from adding uncertainty, but not degree, to the crisp values of subjective judgments.

5 Conclusions & Future Work

In this work, we have presented a novel approach for the fuzzy aggregation of semantic similarity measures. This novel approach can be summarized using the motto **Consensus or Trade-off** what means that atomic semantic similarity measures have to be aggregated without reflecting dissident recommendations in case of a consensus have been reached or using a high degree of trade-off in case a recommendation consensus from atomic measures does not exist. Results show us that **this novel approach is**

able to consistently beat existing approaches based on compensative operators when solving both general purpose and domain specific data sets.

In future, demanding applications where high accuracy of understanding of the user intent is needed, the stakes are high and the users may present adversarial or disruptive characteristics in interacting with systems will require the use of very precise semantic similarity measures. We want to investigate what happens when the amount of linguistic terms for assessing semantic similarity measurement is increased. Additionally, it could be interesting to explore the horizontal aggregation of semantic similarity measures, i.e. the aggregation of single measures of different nature. Positive results in this context could lead to computers to be able to recognize and predict the semantic similarity between text expressions without requiring any kind of human intervention.

Acknowledgments

We would like to thank the reviewers for their time and consideration. This work has been funded by Vertical Model Integration within Regionale Wettbewerbsfaehigkeit OOE 2007-2014 by the European Fund for Regional Development and the State of Upper Austria.

References

[1] Ahlgren, P., Jarneving, B., Rousseau, R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. JASIST (JASIS) 54(6):550-560 (2003).

[2] Barzilay, R., McKeown, K.: Sentence Fusion for Multidocument News Summarization. Computational Linguistics 31(3). 297-328 (2005).

[3] Batet, M., Sanchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J. Biomed. Inform. 44: 118-25 (2010).

[4] Batet, M. Ontology-based semantic clustering. AI Commun. 24(3): 291-292 (2011).

[5] Bickmore, T.W., Giorgino, T. Health dialog systems for patients and consumers. Journal of Biomedical Informatics: 556-571 (2006).

[6] Bollegala, D., Matsuo, Y., Ishizuka, M. A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Trans. Knowl. Data Eng. (TKDE) 23(7):977-990 (2011).

[7] Bollegala, D., Matsuo, Y., Ishizuka, M. Measuring semantic similarity between words using web search engines. WWW 2007: 757-766.

[8] Budanitsky, A., Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32(1): 13-47 (2006).

[9] Chaves-Gonzalez, J.M., Martinez-Gil, J. Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowl.-Based Syst. 37: 62-69 (2013).

[10] Cilibrasi, R., Vitanyi, P. M. B. The Google Similarity Distance. IEEE Trans. Knowl. Data Eng. 19(3): 370-383 (2007).

[11] Chen, B., Foster, G.F., Kuhn, R. Bilingual Sense Similarity for Statistical Machine Translation. ACL 2010:834-843.

[12] Couto, F.M., Silva, M.J., Coutinho, P. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. CIKM 2005:343-344.

[13] Do H.H., Rahm, E. COMA - A System for Flexible Combination of Schema Matching Approaches. VLDB 2002: 610-621.

[14] Erozel, G., Cicekli, N.K., Cicekli, I. Natural language querying for video databases. Inf. Sci. 178(12): 2534-2552 (2008).

[15] Grabisch, M., Marichal, J.L., Mesiar, R., Pap, E. Aggregation functions: Construction methods, conjunctive, disjunctive and mixed classes. Inf. Sci. 181(1): 23-43 (2011).

[16] Grabisch, M. Fuzzy integral for classification and feature extraction. Fuzzy Measures and Integrals: Theory and Applications, 415-434 (2000).

[17] Jiang, J.J., Conrath, D.W. Semantic similarity based on corpus statistics and lexical taxonomy. ROCLING 1997: 19-33.

[18] Lamontagne, L., Lapalme, G. Textual Reuse for Email Response. ECCBR 2004: 242-256.

[19] Leacock, C., Chodorow, M. Combining Local Context and WordNet Similarity for Word Sense Identification, WordNet: An Electronic Lexical Database. 1998. MIT Press.

[20] Martinez-Gil, J., Aldana-Montes, J.F. Semantic similarity measurement using historical google search patterns. Information Systems Frontiers 15(3): 399-410 (2013).

[21] Martinez-Gil, J., Navas-Delgado, I., Aldana-Montes, J.F. MaF: An Ontology Matching Framework. J. UCS 18(2): 194-217 (2012).

[22] Martinez-Gil, J., Aldana-Montes, J.F. Smart Combination of Web Measures for Solving Semantic Similarity Problems. Online Information Review 36(5): 724-738 (2012).

[23] Martinez-Gil, J., Aldana-Montes, J.F. Evaluation of two heuristic approaches to solve the ontology meta-matching problem. Knowl. Inf. Syst. 26(2): 225-247 (2011).

[24] Martinez-Gil, J., Aldana-Montes, J.F. Reverse ontology matching. SIGMOD Record 39(4): 5-11 (2010).

[25] Moschitti, A., Quarteroni. S. Kernels on Linguistic Structures for Answer Extraction. ACL (Short Papers) 2008: 113-116.

[26] Miller, G.A., Charles W.G. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1-28 (1991).

[27] O'Shea, J., Bandar, Z., Crockett, K.A., McLean, D. Benchmarking short text semantic similarity. IJIIDS 4(2): 103-120 (2010)

[28] Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40(3): 288-299 (2007)

[29] Petrakis, E.G.M., Varelas, G., Hliaoutakis, A., Raftopoulou, P. X-similarity: computing semantic similarity between concepts from different ontologies. J. Digit. Inf. Manage. 233-237 (2003).

[30] Pirro, G. A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. 68(11): 1289-1308 (2009).

[31] Punuru, J., Chen, J. Learning non-taxonomical semantic relations from domain texts. J. Intell. Inf. Syst. 38(1): 191-207 (2012).

[32] Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. J. Artif. Intell. Res. (JAIR) 11: 95-130 (1999).

[33] Sanchez, D., Batet, M., Isern, D. Ontology-based information content computation. Knowl.-Based Syst. 24(2): 297-303 (2011).

[34] Tversky, A. Features of Similarity. Psychological Reviews 84 (4): 327352 (1977).

[35] Yager, R.R. Time Series Smoothing and OWA Aggregation. IEEE T. Fuzzy Systems (TFS) 16(4):994-1007 (2008).

Biography

Jorge Martinez-Gil is a Spanish-born computer scientist working in the Knowledge Engineering field. He got his PhD in Computer Science from University of Malaga in 2010. He has hold a number of research position across some European countries (Austria, Germany, Spain). He currently holds a Team Leader position within the group of Knowledge Representation and Semantics from the Software Competence Center Hagenberg (Austria) where he is involved in several applied and fundamental research projects related to knowledge-based technologies. Dr. Martinez-Gil has authored many scientific papers, including those published in prestigious journals like SIGMOD Record, Knowledge and Information Systems, Information Systems Frontiers, Knowledge-Based Systems, Artificial Intelligence Review, Knowledge Engineering Review, Online Information Review, Journal of Universal Computer Science, Journal of Computer Science and Technology, and so on.