Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain

José M. Chaves-González, Jorge Martínez-Gil

University of Extremadura. Escuela Politécnica, 10003, Cáceres, Spain.

{jm, jorgemar}@unex.es

Abstract

Computing accurately the semantic similarity between terms represents an important challenge in the semantic web field. The problem here is the lack of appropriate dictionaries for specific and dynamic domains, such as the biomedical, financial or any other particular field. In this article we propose a new approach which uses different existing semantic similarity methods to obtain precise results which are very close to human judgments in the biomedical domain. Specifically, we have developed an evolutionary algorithm which uses information provided by different semantic similarity metrics. The results provided by our system have been validated using several medical datasets and different sets of similarity functions. We adopt the Pearson correlation coefficient as measure of the strength of the relation between human ratings of similarity and computation values. The proposed approach obtains the best results of correlation with regard to human judgments in comparison with other relevant similarity functions used in the mentioned domains.

Index terms

Semantic similarity, evolutionary computation, semantic web, synonym recognition, differential evolution

I. Introduction

With the increase of large collections of data resources on the World Wide Web (WWW), the study of web semantic techniques [1] has become one of the most active areas for researchers. The notion of semantic web expresses the aspiration to organize the available resources on basis of semantic information in order to facilitate its efficient processing. Related to this, the study of semantic similarity [2] between text expressions is a very relevant problem for certain applications in some specific fields, such as data integration

[3], query expansion [4] or document classification [5] (among others), because they need semantic similarity computation for working appropriately.

On the other hand, semantic similarity measurements are usually performed using some kind of metrics [6]. The most common of those metrics are semantic similarity measures which are a kind of text based metrics resulting in a similarity or dissimilarity score between two given text strings to be compared. A semantic similarity measure provides a floating point number between 0 (total dissimilarity) and 1 (complete similarity).

Most of the existing works have reached a high level of accuracy when solving datasets containing general purpose terms [7]. The majority of these works describe approaches which use large and updated dictionaries [6]. However, most of them often fail when dealing with expressions not covered by these dictionaries. In this work, we propose a new technique which beats a wide range of semantic similarity measurement methods. This technique consists of using an evolutionary algorithm which smartly combines the information provided by several classical semantic similarity functions. Thus, the evolutionary algorithm works as a high level heuristics which was designed with the purpose of improving the results obtained by using each individual similarity function. Therefore, in order to validate our proposal, we use some datasets belonging to the biological domain, where classical algorithms do not usually get the optimal results.

The rest of this work is organized as follows: Section II elaborates on the work carried out on the problem and approaches handled in this paper. Section III describes the problem and the proposed solution. The methodology followed in all the experiments and the results obtained are presented and discussed in Section IV. Finally, conclusions and future lines of research are explained in the last section.

II. Previous Work

Semantic similarity measurement has traditionally been an active research area in the Natural Language Processing (NLP) field [8]. The reason is that the capability for synonym recognition is a key aspect in human conversations. However, with the quick development of the semantic web, researchers from this field have turned their attention to the possibility to adapt these techniques for making easier the discovery of resources from the WWW [9]. In this way, a user who is looking for cars can obtain results including terms such as automobiles, vehicles, and so on. For this reason, a number of publications which work in the intersection of NLP and semantic web have been developed over the last years [10].

On the other hand, a first approach for measuring semantic similarity between words could consist of computing the Euclidean distance (which is one of the most popular metrics in a lot of cultures) between the words. However, Euclidean distance is not appropriate for all types of data or patterns to be compared. For example, this kind of metrics is not appropriate to compute the distance between word meanings. For this reason, most of the previous works have been focused on designing new semantic similarity measures.

Traditionally, these new semantic similarity measures use some kind of dictionaries in order to compute the degree of similarity between the words being compared. Some examples of these dictionaries are WordNet [6], MeSH (Medical Subject Headings) [7] and UMLS (Unified Medical Language System) [12]. These measures can be classified into three main categories:

- *Path-based Measures* which take into account the position of the terms in a given dictionary. If a word has two or more meanings, then multiple paths may exist between the two words. A problem for this approach is that it relies on the notion that all links in the taxonomy represent uniform distances.

- *Information Content Measures.* According to Pedersen et al. [6] information content measures are based on frequency counts of concepts as found in a corpus of text. Within this kind of measures, higher values are associated with more specific concepts (e.g., pitch fork), while those with lower values are more general (e.g., idea).

- *Feature based Measures* which measure the similarity between terms as a function of their properties or based on their relationships to other similar terms in a dictionary. In general, it is possible to estimate semantic similarity according to the amount of common features. An example of feature could be the concept descriptions retrieved from dictionaries.

The problem of traditional semantic similarity metrics is that there are several fields where it is not easy to find complete and updated dictionaries. We focus on the biomedical domain. Related to the problem of semantic similarity measurement in this domain, several works have been proposed in recent years. For instance, Pirró [7] proposed a new information content measure using the MeSH biomedical ontology. Our study improves the results obtained in this study using a combination of several similarity functions. Experimental evaluations indicated that the proposed metrics improved the existing results for a given benchmark dataset. Nguyen and Al-Mubai [11] also proposed an ontology-based semantic similarity measure and applied it into the biomedical domain. The proposed measure is based on the path length between the concept nodes as well as the depth of the terms in the ontology hierarchy tree. Our results in this case are also closer to human judgment. Finally, Pedersen et al. [12] implemented and evaluated a variety of semantic similarity measures based on ontologies and terminologies found in the Unified Medical Language System (UMLS) obtaining very good results. Our results

are better also in this case because we combine the results from other proposals, and we obtain again more precise results as will be explained in the following sections.

III. Differential Evolution for Synonym Recognition

In this section we describe, on the one hand, the problem we have tackled in this paper and, on the other hand, the details of the solution proposed to solve it. Related to this, we explain both our evolutionary approach and the different methods which give support to the system developed. Our approach is based on the similarity result provided by different atomic similarity functions. The evolutionary algorithm works as a hyper-heuristics which assigns different coefficient values to the results of similarity calculated by the pool of functions which are included in the system. Thus, although all the functions (or metrics) are taken into account to provide the final semantic similarity of a specific term pair, the more similar to the human expert will possibly have at the end of the process the highest coefficients. Fig. 1 shows the working diagram of the proposed approach.

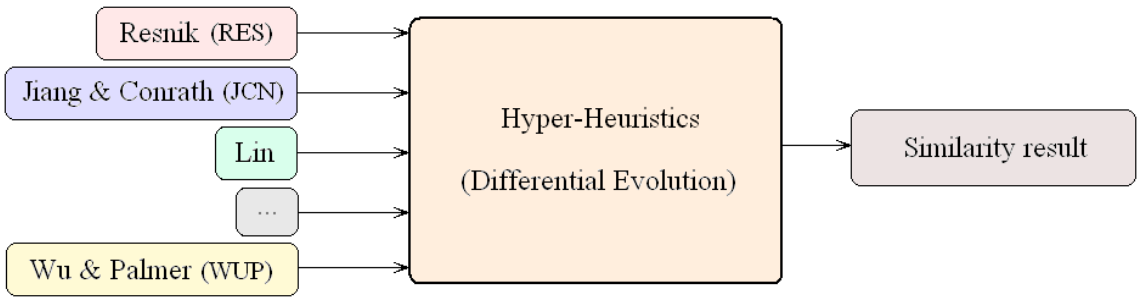

Fig. 1. Working diagram of the proposed approach

The differential evolution (DE) algorithm [13] was chosen among other candidates because after a preliminary study, it was proved that it obtained very competitive results for the problem tackled. The reason is in the way in which the algorithm makes the solution evolve. Due to our system can be considered a hyper-heuristics (HH) which uses the differential evolution (DE) to assign to each similarity function a coefficient which modifies its importance in the whole system, we have called our proposal HH(DE). Differential evolution performs the search of local optima by making small additions and subtractions between the members of its population (see Section III.C), and this feature fits perfectly to the problem, because the algorithm works with the scores provided by the similarity functions (Fig. 1). In fact, the individual is defined as an array of floating point

values, *s*, where $s(f_x)$ is the coefficient value which modifies the result provided by the similarity function f. Fig. 2 illustrates the solution encoded. This floating point value is between MIN and MAX, which is a parameter of the algorithm.

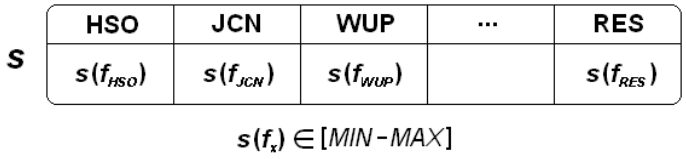

Fig. 2. Individual representation

A. The Synonym Recognition Problem

Given two text expressions *a* and *b*, the problem addressed consists of trying to measure the degree of synonymy between them. The problem of synonym recognition usually extends beyond synonymy and involves semantic similarity measurement. According to Bollegala et al. [14], a certain degree of semantic similarity can be observed not only between synonyms (e.g. lift and elevator), but also between metonyms (e.g. car and wheel), hyponyms (leopard and cat), related words (e.g. blood and hospital) as well as between antonyms (e.g. day and night).

Therefore, we are looking for a computational algorithm which may provide a floating point number indicating automatically the notion of similarity. A value of 0 will stand for not similarity between the (set of) words to be compared, while a value of 1 will indicate that the (set of) words share exactly the same meaning. A more detailed explanation can be found in Section IV.

B. Semantic Similarity Metrics

If we look at the literature, we can find a lot of similarity metrics. In this section we are going to present the similarity metrics that we use as the input for the hyper-heuristic algorithm developed (see Fig. 1).

In Table I the most relevant similarity metrics are summarized according to their classification. The first column indicates the general type of the function. Second column

contains the similarity metrics and a reference in which it is possible to obtain more detailed information. Finally, the third column includes a brief explanation of the metrics.

Table I. Classification of the most relevant similarity metrics.

Type	Similarity function and reference	Brief description
Path based measures	*Path Length* (PATH) [6]	It is inversely proportional to the number of nodes along the shortest path between the words.
	Leacock & Chodorow (LCH) [15]	It can be computed as *-log (length / (2 * D))*, where *length* is the *length* of the shortest path between the two words and *D* is the maximum depth of the taxonomy.
	Wu & Palmer (WUP) [16]	It considers the depths of the two terms in the taxonomies, along with the depth of the LCS (least common subsumer). The formula is: *score* = 2**depth*(*LCS*) / (*depth*(*s1*) + *depth*(*s2*)).
	Hirst & St-Onge (HSO) [17]	It finds lexical chains linking the two word senses.
Information content measures	*Resnik* (RES) [18]	It computes the information content of the LCS.
	Jiang & Conrath (JCN) [19]	It can be computed in the following way: 1 / *IC*(*term1*) + *IC*(*term2*) - 2 * *IC*(*LCS*), where *IC* refers to information content.
	Lin (LIN) [20]	It can be computed as follows: 2 * *IC*(*LCS*) / (*IC*(term1) + IC(term2)).
Feature based measures	*Adapted Lesk* (LESK) [21]	The similarity score is the sum of the squares of the overlap lengths.
	Gloss Vector (vector) [22]	It works by forming second-order co-occurrence vectors from the glosses or WordNet definitions of concepts.
	Gloss Vector (pairwise modified) (vector_pairs) [6]	This metrics forms separate vectors corresponding to each of the adjacent glosses.

The main advantage of the path based measures is that they are very simple to interpret and implement. On the contrary, this kind of measures needs rich taxonomies,

only works with nodes belonging to these taxonomies, and only the relation is-a can be taken into account. The advantage of the Information Content measures is that use empirical information from real corpora. The problem is that only works with nodes (nouns) belonging to these taxonomies, and only is-a relationships can be considered. On the other hand, Feature based measures do not require underlying structures and use implicit knowledge from real large corpora. As a disadvantage, the definitions of the terms can be short, and moreover, the computation can be very intensive in the most of the cases.

C. The Differential Evolution algorithm

Differential Evolution (DE) heuristics [13] is a population based Evolutionary Algorithm (EA) used for function optimization. Due to its simplicity and effectiveness, DE has been applied to a large number of optimization problems in a wide range of domains [23]. Its key idea is based on the generation of new individuals by calculating vector differences between other randomly-selected solutions of the population. The algorithm has been carefully configured and adapted to the problem managed in our study, as will be explained in Section IV. For this reason, among the different variants of the algorithm [24], it was chosen the DE/best/1/bin version, because it provided more competitive results than other versions. The notation (DE/best/1/bin) indicates the way in which the crossover and mutation operators work. Thus, the DE developed includes binomial crossover (bin) and it only uses one (/1/) difference vector in the mutation process of the best solution of the population (best). Algorithm 1 shows an outline of the algorithm developed.

Algorithm 1. Pseudo-code for the DE algorithm

```
1:  generateRandomPopulation (population)
2:  calculateFitness (population)
3:  while (stop condition not reached) do
4:      for (each individual of the population)
5:          selectIndividuals (xTarget, xBest, xInd1, xInd2)
6:          xTrial ← diffMutation (xBest, F, xInd1, xInd2)
7:          xTrial ← binCrossOver (xTarget, xTrial, CrossProb)
8:          calculateFitness (xTrial)
9:          updateIndividual (xTarget, xTrial)
10:     endfor
11: endwhile
12: return bestIndividual (population)
```

The algorithm starts with the random generation of the population (line 1) through the assignment of a random coefficient to each gene of the individual (see Fig. 2). Then, the fitness of each individual will be assigned using the Pearson correlation [25] with the values given by the human expert for all the word pairs of the specific dataset used in the experiment. This value of correlation, *corr*(*X*,*Y*), indicates the quality of the generated solutions and it is defined as expressed in equation 1 (see [25] for a detailed explanation of this equation). The result of this measure is a floating point value between +1 (perfect positive linear correlation) and -1 (perfect negative linear correlation). The result indicates the degree of linear dependence between the variables *X* (the human expert opinion) and *Y* (the solution evaluated). The closer to either -1 or +1, the stronger the correlation between variables and the higher the quality of the solution generated. On the other hand, if the result gets closer to 0, it means that the variables are closer to be uncorrelated, and therefore the solution is considered of poor quality.

$$corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

After the evaluation of the whole population (line 2), each individual will be processed (line 4) to try to improve it. The first thing to do is to select four solutions (line 5). *xTarget* and *xBest* are the solution which is being processed and the best solution found, at that precise moment, respectively. *xInd1* and *xInd2* are two solutions different from *xTarget* and *xBest* chosen randomly. After that, the mutation operation is performed according to the following expression: $xTrial \leftarrow xBest + F \cdot (xInd1 - xInd2)$. This operator generates the *xTrial* individual which will be compared to the *xTarget* after the mutation and crossover operations (line 9). The parameter $F \in [0, 2]$ establishes the mutation which is going to suffer the best solution, *xBest*. After the mutation, the *xTarget* and *xTrial* individuals are crossed using a binary crossover [26] according a certain crossover probability, $crossProb \in [0, 1]$. The obtained individual is evaluated (equation 1) to check its quality (line 8) and finally will be compared with the *xTarget* solution. The best individual, or the best set of coefficients which modifies the metrics used by the system (Fig. 1), will be saved in the *xTarget* position and the other will be discarded. This process is repeated for each individual in the population (line 4) while the stop condition is not satisfied (line 3). In our case, the stop condition is a certain number of generations which is also an algorithm parameter to be configured (see Section IV.A). At the end of the process, the best individual, or set of coefficients (Fig. 2) which provides the best similarity results for the dataset which is being tackled, is returned as final result of our system (line 12).

IV. Experiments and Results

In this section we summarize the main experiments and the results obtained in our study. We have used different similarity metrics and biomedical datasets to test the proposed system.

A. Methodology

All the experiments performed have been carried out under the same conditions: an Intel Xeon 2.33 GHz processor and 8 GB RAM. On the software side, we have used the GCC 4.1.2 compiler on a Scientific Linux 5.3 64 bits OS. Since we are dealing with a stochastic algorithm, we have carried out 100 independent runs for each experiment performed. Results provided in the following subsections are average results of these 100 executions. It is important to point here that the use of the arithmetic mean is a valid statistical measurement because the results obtained follow a normal distribution (a p-value greater than 0.05 for the Shapiro-Wilk test confirms this fact [27]). Moreover, all the results present an extremely low dispersion, since the standard deviation for all the experiments is lower than 10^{-15}, so results can be considered statistically reliable.

In the following subsection we will discuss the results obtained with the different experiments using different similarity metrics and different word datasets, but before doing this, it is necessary to present the parameter setting for the algorithm developed. This configuration is very relevant, since the quality of the results obtained by the final system depends largely on the accuracy of this adjustment. Therefore, we performed a complete and precise parameter study for each parameter. All the results presented in the following subsections have been obtained using the same parameter setting, which is summarized in Table II.

Table II. Optimal parameter setting.

Parameter	Optimal value
Population size	100
Mutation factor, F	0.5
Crossover probability, *crossProb*	0.1
Max generations	100
MIN, MAX	-100, 100

B. Result discussion

In this section we discuss the results obtained in different tests. We have performed two set of experiments. First, we explain the results obtained using our proposed system with different similarity metrics (taken from the WordNet dictionary[1] and from the Pirró study [7]). After that we discuss the results obtained using two different datasets from the biomedical domain [7, 11].

1) Experiments with different metrics

In this subsection we study the results provided by our system using two different set of similarity functions. First, we compare our results with the study published by Pirró [7]. In that study the author proposes a new metrics based on features (P&S). Table III summarizes the results of the study with the same biomedical dataset. All values are normalized in the interval [0, 1].

Table III. Similarity results obtained by our system (last column) compared with other results published.

	Word pair		Human	IC Based			Hybrid	Features	EA
	Word 1	Word 2	Expert	Resnik	Lin	J&C	Li	P&S	HH(DE)
P01	Anemia	Appendicitis	0.031	0.000	0.000	0.190	0.130	0.133	0.116
P02	Otitis Media	Infantile Colic	0.156	0.000	0.000	0.160	0.100	0.000	0.056
P03	Dementia	Atopic Dermatitis	0.060	0.000	0.000	0.290	0.130	0.202	0.165
P04	Bacterial Pneumonia	Malaria	0.156	0.000	0.000	0.030	0.100	0.000	0.024
P05	Osteoporosis	Patent Ductus Arteriosus	0.156	0.000	0.000	0.150	0.000	0.000	0.037
P06	Sequence	Antibacterial Agents	0.155	0.000	0.000	0.270	0.160	0.184	0.159
P07	Acq. Immunno. Syndrome	Congenital Heart Defects	0.060	0.000	0.000	0.070	0.080	0.000	0.030
P08	Meningitis	Tricuspid Atresia	0.031	0.000	0.000	0.190	0.130	0.131	0.115
P09	Sinusitis	Mental Retardation	0.031	0.000	0.000	0.360	0.130	0.117	0.152
P10	Hypertension	Failure	0.500	0.000	0.000	0.210	0.130	0.109	0.112
P11	Hyperlipidemia	Hyperkalemia	0.156	0.331	0.483	0.470	0.510	0.561	0.443
P12	Hypothyroidism	Hyperthyroidism	0.406	0.619	0.726	0.750	0.630	0.718	0.665
P13	Sarcoidosis	Tuberculosis	0.406	0.000	0.000	0.250	0.070	0.169	0.134
P14	Vaccines	Immunity	0.593	0.000	0.000	0.520	0.000	0.344	0.251
P15	Asthma	Pneumonia	0.375	0.517	0.790	0.870	0.520	0.749	0.627
P16	Diabetic Nephropathy	Diabetes Mellitus	0.500	0.612	0.759	0.790	0.770	0.741	0.696
P17	Lactose Intolerance	Irritable Bowel Syndrome	0.468	0.468	0.468	0.470	0.360	0.468	0.451
P18	Urinary Tract Infection	Pyelonephritis	0.656	0.470	0.588	0.670	0.420	0.604	0.533
P19	Neonatal Jaundice	Sepsis	0.187	0.000	0.000	0.190	0.160	0.000	0.073
P20	Anemia	Deficiency Anemia	0.437	0.601	0.720	0.790	0.360	0.712	0.622
P21	Psychology	Cognitive Science	0.593	0.627	0.770	0.810	0.800	0.751	0.714
P22	Adenovirus	Rotavirus	0.437	0.267	0.332	0.450	0.350	0.398	0.358
P23	Migraine	Headache	0.718	0.229	0.243	0.370	0.170	0.269	0.266
P24	Myocardial Ischemia	Myocardial Infarction	0.750	0.595	0.918	0.890	0.800	0.830	0.713
P25	Hepatitis B	Hepatitis C	0.562	0.649	0.823	0.860	0.660	0.790	0.715
P26	Carcinoma	Neoplasm	0.750	0.246	0.626	0.850	0.450	0.651	0.488
P27	Pulmonary Stenosis	Aortic Stenosis	0.531	0.658	0.781	0.810	0.660	0.763	0.707
P28	Failure to Thrive	Malnutrition	0.625	0.000	0.000	0.180	0.130	0.126	0.111
P29	Breast Feeding	Lactation	0.843	0.000	0.000	0.040	0.080	0.029	0.033
P30	Antibiotics	Antibacterial Agents	0.937	1.000	1.000	1.000	0.990	1.000	1.000
P31	Seizures	Convulsions	0.843	0.880	1.000	0.900	0.810	0.990	0.887
P32	Pain	Ache	0.875	0.861	1.000	1.000	0.990	0.954	0.920
P33	Malnutrition	Nutritional Deficiency	0.875	0.622	1.000	1.000	0.980	0.874	0.780
P34	Measles	Rubeola	0.906	0.924	1.000	1.000	0.990	1.000	0.965
P35	Chicken Pox	Varicella	0.968	1.000	1.000	1.000	0.990	1.000	1.000
P36	Down Syndrome	Trisomy 21	0.875	1.000	1.000	1.000	0.990	1.000	1.000

[1] http://wordnet.princeton.edu/

The first column identifies de word pair. Then the word pair which is under evaluation is presented (columns 2 and 3). The fourth column corresponds with the value of correlation with regard to human judgment. After that, all the computational methods appear classified in different methodologies. Our results are shown in the last column. As described previously, they are highly reliable since they are the mean result of 100 independent executions, with a very low standard deviation.

Table IV presents the values of Pearson correlation between the methods which appeared in Table III, and which were taken from [7], and the human expert value. As can be observed, our approach (HH(DE)) provides the best results of the study.

Table IV. Pearson correlation between computational methods and human judgments.

Similarity function	Correlation
Resnik	0.721
Lin	0.718
J&C	0.718
Li	0.707
P&S	0.725
HH(DE)	**0.732**

Table V. Similarity results obtained by our system (last column) compared with the results obtained from WordNet.

Word pair	Human Expert	HSO	JCN	WUP	PATH	LIN	LESK	RES	LCH	vector_pairs	vector	HH(DE)
P01	0.031	0.250	0.000	0.842	0.250	0.000	0.004	0.000	0.624	0.010	0.183	0.139
P02	0.156	0.188	0.000	0.800	0.200	0.000	0.002	0.000	0.564	0.007	0.211	0.169
P03	0.06	0.000	0.000	0.480	0.071	0.000	0.002	0.000	0.285	0.013	0.111	0.115
P04	0.156	0.125	0.092	0.720	0.250	0.524	0.010	0.000	0.436	0.080	0.326	0.113
P05	0.156	0.000	0.000	0.174	0.050	0.000	0.002	0.000	0.188	0.018	0.081	0.034
P06	0.155	0.000	0.058	0.300	0.077	0.060	0.010	0.000	0.305	0.022	0.152	0.073
P07	0.06	0.000	0.052	0.375	0.091	0.075	0.028	0.000	0.350	0.051	0.397	0.195
P08	0.031	0.000	0.000	0.609	0.100	0.000	0.002	0.000	0.376	0.011	0.121	0.134
P09	0.031	0.000	0.000	0.556	0.111	0.000	0.003	0.000	0.404	0.003	0.057	0.083
P10	0.5	0.250	0.000	0.842	0.250	0.000	0.007	0.000	0.624	0.018	0.457	0.281
P11	0.156	0.313	0.000	0.889	0.333	0.000	0.078	0.331	0.702	0.195	0.727	0.472
P12	0.406	1.000	0.000	0.900	0.333	0.000	0.019	0.619	0.702	0.221	0.097	0.199
P13	0.406	0.000	0.000	0.720	0.125	0.000	0.007	0.000	0.436	0.070	0.222	0.165
P14	0.593	0.000	0.048	0.267	0.083	0.000	0.009	0.000	0.326	0.016	0.251	0.125
P15	0.375	0.313	0.000	0.923	0.333	0.000	0.013	0.517	0.702	0.227	0.375	0.358
P16	0.5	1.000	0.044	0.182	0.053	0.000	0.105	0.612	0.202	0.041	0.396	0.435
P17	0.468	0.000	0.000	0.250	0.053	0.000	0.001	0.468	0.202	0.034	0.108	0.298
P18	0.656	0.250	0.000	0.963	0.500	0.000	0.050	0.470	0.812	0.062	0.591	0.536
P19	0.187	0.125	0.059	0.400	0.077	0.221	0.011	0.000	0.305	0.043	0.112	0.029
P20	0.437	0.000	0.108	0.571	0.100	0.471	0.051	0.601	0.376	0.074	0.329	0.450
P21	0.593	0.375	0.000	0.900	0.333	0.000	0.153	0.627	0.702	0.167	0.515	0.539
P22	0.437	0.250	0.000	0.842	0.250	0.000	0.029	0.267	0.624	0.042	0.093	0.212
P23	0.718	0.250	0.000	0.957	0.500	0.000	0.428	0.229	0.812	0.020	0.612	0.504
P24	0.75	0.000	0.000	0.720	0.125	0.000	0.038	0.595	0.436	0.034	0.116	0.446
P25	0.562	0.313	0.000	0.933	0.333	0.000	0.074	0.649	0.702	0.359	0.583	0.463
P26	0.75	0.313	0.000	0.889	0.250	0.000	0.228	0.246	0.624	0.134	0.480	0.385
P27	0.531	0.313	0.000	0.917	0.333	0.000	0.100	0.658	0.702	0.438	0.833	0.550
P28	0.625	0.000	0.000	0.636	0.111	0.000	0.029	0.000	0.404	0.125	0.309	0.164
P29	0.843	0.250	0.000	0.857	0.250	0.000	0.017	0.000	0.624	0.218	0.849	0.366
P30	0.937	0.250	1.000	0.941	0.500	1.000	0.661	1.000	0.812	0.119	0.769	0.894
P31	0.843	0.313	0.455	0.952	0.500	0.897	0.098	0.880	0.812	0.197	0.628	0.710
P32	0.875	0.313	0.402	0.947	0.500	0.861	0.152	0.861	0.812	0.050	0.419	0.562
P33	0.875	0.000	0.000	0.571	0.100	0.000	0.040	0.622	0.376	0.022	0.322	0.554
P34	0.906	1.000	0.000	1.000	1.000	0.000	1.000	0.924	1.000	0.333	1.000	0.791
P35	0.968	0.250	0.000	0.966	0.500	0.000	0.752	1.000	0.812	0.055	0.720	1.000
P36	0.875	1.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	0.500	1.000	0.720

Furthermore, we checked our approach using a different set of similarity functions. In this second case we used the WordNet::similarity resources [6]. The word dataset used was exactly the same than in the previous set of experiments. All the metrics available in WordNet have been included in our study. Table V presents the results obtained with this second set of metrics.

All the metrics used have been previously explained in Section III.B (Table I). Results obtained by our system appear in the last column. As was mentioned before, our results are statistically robust since they are the mean result of 100 independent executions, with a very low standard deviation. Although there are functions which do not obtain good correlation results (Table VI), such as HSO (0.332), JCN (0.237), LIN (0.218) or vector_pairs (0.333), they have been included in our study, since they provide significant contribution to the HH(DE) developed. This can be checked through Table VII, where we can observe the correlation value obtained by our system reducing the similarity functions included. As can be observed, although the global correlation of a particular function is not very good, that particular function is important in our system, since our hyper-heuristics adapts the importance of that function by varying its coefficient.

Table VI. Pearson correlation between computational methods calculated in WordNet::similarity and human judgments.

Metrics	Correlation
HSO	0.332
JCN	0.237
WUP	0.490
PATH	0.517
LIN	0.218
LESK	0.517
RES	0.721
LCH	0.553
vector_pairs	0.333
vector	0.593
HH(DE)	**0.809**

As can be observed in Table VI, our proposal improves significantly all the rest of metrics. In fact, the correlation value reached (0.809) is even better than the result obtained in the previous set of experiments (0.732, Table IV), although the correlation obtained by the individual similarity functions was of more quality. Therefore, we can conclude that the more similarity functions used by our approach, the more quality in the results obtained.

Table VII. Pearson correlation values for different number of similarity functions. HH(DE) uses the similarity functions of more quality in each case.

Similarity functions used by HH(DE)	Correlation
10	**0.809**
9	0.771
8	0.769
7	0.768
6	0.701
5	0.692
4	0.692
3	0.678
2	0.658
1	0.642

2) Experiments with different datasets

In this subsection we study the results provided by our system using other biomedical dataset [11]. The configuration of our system is exactly the same, therefore, we can say that the parametrical setting of our HH(DE) is consistent for the two datasets checked. To our best knowledge, there are no works in which more datasets from this specific domain have been used, so we cannot perform more comparisons. Table VIII presents the word pairs of the dataset and the expert value provided by human experts. As in the previous case, all values are normalized in the interval [0, 1].

Table VIII. Word dataset used in the second set of experiments.

Word pair	Word 1	Word 2	Human Expert
WP01	Renal failure	Kidney failure	1
WP02	Heart	Myocardium	0.75
WP03	Stroke	Infarct	0.7
WP04	Abortion	Miscarriage	0.825
WP05	Delusion	Schizophrenia	0.55
WP06	Congestive heart failure	Pulmonary edema	0.35
WP07	Metastasis	Adenocarcinoma	0.45
WP08	Calcification	Stenosis	0.5
WP09	Diarrhea	Stomach cramps	0.325
WP10	Mitral stenosis	Atrial fibrillation	0.325
WP11	Chronic obstructive pulmonary disease	Lung infiltrates	0.475
WP12	Rheumatoid arthritis	Lupus	0.275
WP13	Brain tumor	Intracranial hemorrhage	0.325
WP14	Carpel tunnel syndrome	Osteoarthritis	0.275
WP15	Diabetes mellitus	Hypertension	0.25
WP16	Acne	Syringe	0.25
WP17	Antibiotic	Allergy	0.3
WP18	Cortisone	Total knee replacement	0.25
WP19	Pulmonary embolus	Myocardial infarction	0.3
WP20	Pulmonary fibrosis	Lung cancer	0.35
WP21	Cholangiocarcinoma	Colonoscopy	0.25
WP22	Lymphoid hyperplasia	Laryngeal cancer	0.25
WP23	Multiple sclerosis	Psychosis	0.25
WP24	Appendicitis	Osteoporosis	0.25
WP25	Rectal polyp	Aorta	0.25
WP26	Xerostomia	Alcoholic cirrhosis	0.25
WP27	Peptic ulcer disease	Myopia	0.25
WP28	Depression	Cellulites	0.25
WP29	Varicose vein	Entire knee meniscus	0.25
WP30	Hyperlidpidemia	Metastasis	0.25

Our results are shown in Table IX. The two first columns are taken from Table VIII and identify the word pair and the human expert value. The following 10 columns corresponds with the values obtained from the WordNet similarity tool [6], and the last column contains the result obtained by our HH(DE). As previously, this results are statistically confident because they are the mean result of 100 independent executions, with a very low standard deviation (lower than 10^{-10}).

Table IX. Similarity results obtained by our system (last column) compared with the results obtained from WordNet for the second dataset.

Word pair	Human Expert	HSO	JCN	WUP	PATH	LIN	LESK	RES	LCH	vector _pairs	vector	HH(DE)
WP01	1	1.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	0.010	0.183	1.000
WP02	0.75	0.313	0.078	0.600	0.111	0.370	0.210	0.320	0.404	0.007	0.211	0.517
WP03	0.7	0.000	0.055	0.333	0.077	0.079	0.060	0.066	0.305	0.013	0.111	0.094
WP04	0.825	1.000	0.000	1.000	1.000	0.000	0.195	0.907	1.000	0.080	0.326	0.742
WP05	0.55	0.188	0.000	0.778	0.200	0.000	0.114	0.469	0.564	0.018	0.081	0.193
WP06	0.35	0.000	0.000	0.556	0.111	0.000	0.047	0.269	0.404	0.022	0.152	0.019
WP07	0.45	0.000	0.000	0.174	0.050	0.000	0.018	0.000	0.188	0.036	0.327	0.050
WP08	0.5	0.000	0.000	0.556	0.111	0.000	0.027	0.269	0.404	0.051	0.397	0.030
WP09	0.325	0.000	0.057	0.333	0.077	0.000	0.074	0.066	0.305	0.011	0.121	0.123
WP10	0.325	0.188	0.000	0.833	0.200	0.000	0.022	1.000	0.564	0.003	0.057	0.128
WP11	0.475	0.000	0.000	0.500	0.200	0.000	0.008	0.297	0.298	0.018	0.457	0.030
WP12	0.275	0.188	0.000	0.846	0.200	0.000	0.114	0.582	0.564	0.195	0.727	0.079
WP13	0.325	0.000	0.098	0.750	0.143	0.540	0.043	0.508	0.472	0.221	0.097	0.056
WP14	0.275	0.000	0.000	0.160	0.046	0.000	0.029	0.000	0.162	0.070	0.222	0.055
WP15	0.25	0.000	0.000	0.560	0.083	0.000	0.095	0.477	0.326	0.016	0.251	0.124
WP16	0.25	0.000	0.043	0.167	0.048	0.000	0.036	0.000	0.175	0.227	0.375	0.023
WP17	0.3	0.000	0.000	0.200	0.059	0.000	0.077	0.000	0.232	0.041	0.396	0.125
WP18	0.25	0.000	0.000	0.300	0.067	0.000	0.040	0.052	0.266	0.034	0.108	0.018
WP19	0.3	0.000	0.000	0.300	0.067	0.000	0.037	0.066	0.266	0.062	0.591	0.066
WP20	0.35	0.000	0.000	0.667	0.100	0.000	0.022	0.508	0.376	0.043	0.112	0.007
WP21	0.25	0.000	0.000	0.222	0.046	0.000	0.066	0.066	0.162	0.500	1.000	0.298
WP22	0.25	0.000	0.104	0.583	0.091	0.540	0.045	0.477	0.350	0.074	0.329	0.029
WP23	0.25	0.000	0.000	0.632	0.125	0.000	0.052	0.352	0.436	0.167	0.515	0.029
WP24	0.25	0.000	0.000	0.286	0.063	0.000	0.012	0.066	0.248	0.042	0.093	0.010
WP25	0.25	0.000	0.000	0.261	0.056	0.000	0.033	0.052	0.216	0.020	0.612	0.018
WP26	0.25	0.000	0.000	0.571	0.100	0.000	0.014	0.352	0.376	0.034	0.116	0.005
WP27	0.25	0.000	0.000	0.692	0.111	0.000	0.022	0.508	0.404	0.359	0.583	0.191
WP28	0.25	0.000	0.000	0.375	0.091	0.000	0.025	0.052	0.350	0.134	0.480	0.081
WP29	0.25	0.000	0.000	0.546	0.091	0.000	0.021	0.320	0.350	0.438	0.833	0.221
WP30	0.25	0.000	0.000	0.250	0.077	0.000	0.048	0.000	0.305	0.125	0.309	0.170

As occurred for the other dataset, our results clearly improve the result provided by the rest of metrics used. In fact, we can observe how the HSO metrics obtained quite poor results for the previous dataset (0.332, Table VI) and in this case obtain quite nice results (0.701, Table X) using in both cases WordNet. Therefore, results obtained by our approach are more reliable for different word datasets than other similarity functions.

Finally, Table XI summarizes all the results that we found related with this second dataset. As we can observe, our approach improves any other similarity function applied over the same word dataset.

Table X. Correlation between computational methods and human judgments for the second dataset.

Metrics	Correlation
HSO	0.701
JCN	0.111
WUP	0.483
PATH	0.753
LIN	0.077
LESK	0.712
RES	0.106
LCH	0.687
vector_pairs	-0.351
vector	-0.289
HH(DE)	**0.885**

Table XI. Comparison of the correlation value obtained by several approaches. The reference of the work is included.

Similarity function (metric)	Correlation
Vector [12]	0.76
LIN [12]	0.69
J&C[12]	0.55
RES[12]	0.55
Path[12]	0.48
L&C[12]	0.47
PATH [11]	0.818
L&C[11]	0.833
W&P[11]	0.778
C&K[11]	0.702
Proposed metrics in [11]	0.836
HH(DE)	**0.885**

V. Conclusions and Future Work

In this work, we have presented a novel approach that is able to beat existing similarity functions when dealing with datasets from the biomedical domain. The novelty of our work consists of using other similarity functions as black boxes which are combined in a smart way. This fact produces an important profit for our HH(DE), since it takes important features extracted from the different similarity functions.

We think that our contribution is twofold: to the best of our knowledge, it is the first time that an evolutionary algorithm is used to tackle this problem. And, as our results show, our DE approach obtains very competitive correlation values (see Section IV). In fact, compared with other relevant works published in the bibliography, our approach obtains the highest similarity scores until now.

As future work, we propose to explore further possibilities for synonym recognition in other domains. We are especially interested in areas where good synonym dictionaries do not exist, since we assume that any kind of computational algorithm cannot overcome human knowledge. Moreover, we think to devote efforts to improve our fitness function, so that it can be independent of the domain. Our final goal is to obtain more powerful mechanisms for synonym recognition which can help to reach a real semantic web.

References

[1] N. Shadbolt, T. Berners-Lee, W. Hall, "The Semantic Web Revisited," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96-101, 2006.

[2] A. Budanitsky, G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13-47, 2006.

[3] A.Y. Halevy, A. Rajaraman, J.J. Ordille, "Data Integration: The Teenage Years," *VLDB*, pp. 9-16, 2006.

[4] F.A. Grootjen, T.P. van der Weide. "Conceptual query expansion," *Data Knowledge Engineering*, vol. 56, no. 2, pp. 174-193, 2006.

[5] J. Hu, R.S. Kashi, G.T. Wilfong, "Comparison and Classification of Documents Based on Layout Similarity," *Information Retrieval*, vol. 2, no. 2/3, pp. 227-243, 2000.

[6] T. Pedersen, S. Patwardhan, J. Michelizzi, "Word-Net::Similarity - Measuring the Relatedness of Concepts," *Association for the Advancement of Artificial Intelligence*, pp. 1024-1025, 2004.

[7] G. Pirro, "A semantic similarity metric combining features and intrinsic information content," *Data Knowl. Eng.*, vol. 68, no. 11, pp. 1289-1308, 2009.

[8] C.D. Manning, H. Schütze, "Foundations of Statistical Natural Language Processing," *MIT Press*, Cambridge, Massachusetts, 1999.

[9] E. Kaufmann, A. Bernstein, "Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases," *J. Web Sem.*, vol. 8, no. 3, pp. 377-393, 2010.

[10] A. Java, S. Nirenburg, M. McShane, T.W. Finin, J. English, A. Joshi, "Using a Natural Language Understanding System to Generate Semantic Web Content," *Int. J. Semantic Web Inf. Syst.*, vol. 3, no. 4, pp. 50-74, 2007.

[11] H. Al-Mubaid, H.A. Nguyen, "Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 39, no. 4, pp. 389-398, 2009.

[12] T. Pedersen, S. Pakhomov, S. Patwardhan, C.G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of Biomedical Informatics*, vol. 40, no. 3, pp. 288-299, 2007.

[13] R. Storn, K. Price, "Differential Evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces," *TR-95-012*, International Computer Science Institute, Berkeley, 1995.

[14] D. Bollegala, Y. Matsuo, M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *Proceedings of the World Wide Web Conference*, pp. 757-766, 2007.

[15] C. Leacock, M. Chodorow, G.A. Miller, "Using Corpus Statistics and WordNet Relations for Sense Identification," *Computational Linguistics*, vol. 24, no. 1, pp. 147-165, 1998.

[16] Z. Wu, M. Palmer, "Verb semantics and lexical selection," In *Assoc Comput Linguist Proc.*, pp. 133-138, Las Cruces, NM, USA, 1994.

[17] G. Hirst, D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," In C. Fellbaum, ed., WordNet: An electronic lexical database, *MIT Press*, 1998.

[18] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *IJCAI*, pp. 448-453, 1995.

[19] D. Conrath, J. Jiang, "Semantic similarity based on corpus statistics and lexical taxonomy," In *Comp Linguist Proc.*, pp. 19-33, Taiwan, 1997.

[20] D. Lin, "An information-theoretic definition of similarity," In *Intl Conf ML Proc.*, pp. 296-304, San Francisco, CA, USA, 1998.

[21] M. Lesk, "Information in Data: Using the Oxford English Dictionary on a Computer," *SIGIR Forum*, vol. 20, no. 1-4, pp. 18-21, 1986.

[22] S. Banerjee, T. Pedersen, "Extended Gloss Overlaps as a Measure of Semantic Relatedness," *IJCAI*, pp. 805-810, 2003.

[23] S. Das, P.N. Suganthan, "Differential Evolution: A Survey of the State-of-the-Art," *IEEE Transaction on Evolutionary Computation*, vol. 15, no. 1, pp. 4-31, 2011.

[24] E. Mezura-Montes, J. Velázquez-Reyes, C.A. Coello-Coello, "A comparative study of differential evolution variants for global optimization," In *Proceedings of the 8th annual conference on Genetic and evolutionary computation (GECCO '06)*. ACM, New York, NY, USA, pp. 485-492, 2006.

[25] J.S. Simonoff, “Smoothing methods in statistics,” *Springer*, 1996.

[26] D. Zaharie, “A Comparative Analysis of Crossover Variants in Differential Evolution,” In *Proceedings of the International Multiconference on Computer Science and Information Technology*, Wisla, Poland, pp. 171-181, 2007.

[27] J. Demsar, “Statistical comparison of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.