Int J Software Informatics, Vol.4, No.1, March 2010, pp. 39–52 E-mail: ijsi@iscas.ac.cn
International Journal of Software and Informatics, ISSN 1673-7288 http://www.ijsi.org
 Tel: +86-10-62661040

Web-Based Semantic Similarity: An Evaluation in the Biomedical Domain*

David Sánchez, Montserrat Batet, and Aida Valls

(Intelligent Technologies for Advanced Knowledge Acquisition Research Group, Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Spain)

Abstract Computation of semantic similarity between concepts is a very common problem in many language related tasks and knowledge domains. In the biomedical field, several approaches have been developed to deal with this issue by exploiting the structured knowledge available in domain ontologies (such as SNOMED-CT or MeSH) and specific, closed and reliable corpora (such as clinical data). However, in recent years, the enormous growth of the Web has motivated researchers to start using it as the corpus to assist semantic analysis of language. This paper proposes and evaluates the use of the Web as background corpus for measuring the similarity of biomedical concepts. Several ontology-based similarity measures have been studied and tested, using a benchmark composed by biomedical terms, comparing the results obtained when applying them to the Web against approaches in which specific clinical data were used. Results show that the similarity values obtained from the Web for ontology-based measures are at least and even more reliable than those obtained from specific clinical data, showing the suitability of the Web as information corpus for the biomedical domain.

Key words: semantic similarity; ontologies; information content; Web; biomedicine; UMLS; SNOMED

Sánchez D, Batet M, Valls A. Web-Based semantic similarity: An evaluation in the biomedical domain. *Int J Software Informatics*, **2010, 4(1): 39–52. http://www.ijsi.org/1673-7288/4/i43.htm**

1 Introduction

The computation of the *semantic similarity* between concepts has been a very active trend in computational linguistics. It gives a clue to quantify how words extracted from documents or textual descriptions are alike. Semantically, similarity is usually based on taxonomical relations between concepts. For example, *bronchitis* and *flu* are similar because both are disorders of the respiratory system. However, words can be related in other non-taxonomical ways (e.g. *diuretics* help to treat *hypertension*). In those more general cases, we talk about *semantic relatedness*.

* This work is sponsored by the University Rovira i Virgili (2009AIRE-04), the Spanish Ministry of Science and Innovation (DAMASK project, Data mining algorithms with semantic knowledge, TIN2009-11005) and the Spanish Government (PlanE, Spanish Economy and Employment Stimulation Plan). Montserrat Batet is also supported by a research grant provided by Universitat Rovira i Virgili
Corresponding author: David Sánchez, Email: david.sanchez@urv.cat
Received 2009-07-15; revised 2010-03-20; accepted 2010-03-31.

From a domain independent point of view, the assessment of similarity has many direct applications such as, word-sense disambiguation[38], document categorization or clustering[1,10], word spelling correction[3], automatic language translation[10], information retrieval[14,21] and ontology learning[34].

In this last case, for example, the discovery of taxonomically similar or non-taxonomically related terms to concepts already existing in an ontology (by analysing domain corpora), enables enriching the ontology with new classes and relations in an automated fashion. Due to the manual knowledge representation bottleneck, approaches which may aid to the development of ontologies are very convenient for the Semantic Web.

In the biomedical domain, semantic similarity measures can improve the performance of Information Retrieval tasks[28]. Concretely, similarity computation enables obtaining semantically equivalent words which are useful to reformulate or to expand user queries to multiple formulations. In this manner, IR recall can be improved by retrieving a wider corpus of results which would remain hidden due to the use of strict query-matching search algorithms. This is specially interesting in biomedicine due to the proliferation of domain terminology with different lexicalizations, synonyms, acronyms or abbreviations referring to the same term. Authors have also applied them to discover similar protein sequences[22] or to the automatic indexing and retrieval of biomedical documents (e.g. in the PubMed digital library)[37].

In general, similarity assessment is based on the estimation of semantic evidence observed in one or several knowledge or information sources. So, background data or knowledge is needed in order to measure the degree of similarity between a pair of concepts.

Domain-independent approaches[16,20,29,38] typically rely on WordNet[13], which is a freely available lexical database that represents an ontology of more than 100,000 general English concepts, and which contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (synsets) each expressing a distinct concept, and/or general corpora like SemCor[21], a semantically tagged text repository consisting of 100 passages from the Brown Corpus. A domain-independent scenario is typically characterized by its high amount of ambiguous words, typically affected by polysemy and synonymy. In those cases, the more background knowledge is available (i.e. textual corpus, dictionaries, ontologies, etc.) and the more preprocessing of the corpus data (e.g. manual tagging, disambiguation, etc.), the better the estimation will be Ref.[18].

However, for specialized domains such as biomedicine, words are much less polysemic and unequivocally refer to the corresponding concept. Thus, ambiguity is reduced and the conclusions that could be extracted from the available data may be more accurate.

In the past, some classical similarity computation measures have been adapted to the biomedical domain, exploiting medical ontologies (such as UMLS or MeSH) and/or clinical data sources in order to extract the semantic evidence in which they base the similarity assessment. The general motivation is that the lack of domain coverage of typically exploited domain-independent sources (such as the Brown Corpus or WordNet) makes them ineffective in domain specific tasks[28].

However, the use of a domain dependant corpus introduces some problems: i)

the preparation of the input data for each domain in a format which can be exploited (i.e. data pre-processing or filtering is typically required), ii) data sparseness of scarce concepts if the corpus does not provide enough semantic evidences to extract accurate conclusions, and iii) the availability of enough domain data (this is especially critical in the medical domain, due to the privacy of clinical data). So, even though a domain-dependant approach may lead to more accurate results, the dependency on the domain knowledge and data availability hampers the real applicability of the similarity measures.

The situation has changed in the recent years with the enormous development of the Web. Nowadays, the Web is the biggest electronic repository available[5]. Web data, understood as individual observations in web resources, may seem unreliable due to its uncontrolled publication and "dirty" form. However, considering the Web as a global-scale social source, it has been demonstrated that the amount and heterogeneity of the information available is so high that it can approximate the real distribution of information at a social scale[10]. In fact, some authors[10,36] have exploited web information distribution in order to evaluate word relatedness in an unsupervised fashion (i.e. no domain knowledge is employed). However, their performance is still far[15] from the supervised (ontology-based) approaches studied in this paper.

Following these premises, our hypothesis is that the amount of information related to the biomedical domain (and in general to any concrete domain) contained in the Web is enough to obtain similarity assessments as robust (or even more) as those extracted from a reliable, pre-processed and domain specific corpus (i.e. clinical data, in the case of the biomedical domain). In order to prove it, we have studied the performance of classical ontology-based similarity measures applied to rank the similarity between concepts of the biomedical domain; in our experiments the Web is used as the source from which to perform the semantic assessment, in comparison to other approaches using reliable domain specific data.

This paper presents the analysis and results of this study. In section 2 we will present ontology-based similarity computation paradigms and the way in which they have been used in the past when dealing with biomedical concepts. In sections 3 and 4, we will study and adapt some corpus-based measures to the Web environment (particularly in the way in which word statistics are computed). Then, in section 5, we will evaluate the Web-based measures using a standard benchmark composed by 30 medical terms whose similarity has been assessed by expert physicians of the Mayo Clinic[28] and compare the results against previous approaches evaluating the same measures but using only domain-specific data. The final section will present the conclusions of this study.

2 Semantic Similarity Estimation Paradigms

In the literature, there exist several semantic similarity estimation paradigms according to the techniques employed and the knowledge exploited to perform the assessment.

First, there are unsupervised approaches (i.e. they do not exploit knowledge structures like ontologies) in which semantics are estimated from the information distribution of terms (instead of concepts) in a given corpus[12,18]. Statistical analysis and shallow linguistic parsing are used to measure the degree of co-occurrence

between terms which is used as an estimation of similarity[19]. These are collocation-based measures following the premise that term co-occurrence is an evidence of their relatedness. These measures need a corpus as general as possible (like the Web) in order to estimate social-scale word usage. Particularly, authors have exploited the Web by estimating term collocation probabilities at a social scale from the *web hit count* provided by a Web search engine when querying each or both terms with respect to the total amount of web resources. Concretely, Turney[36] adapted the classical *Pointwise Mutual Information*[7] calling it PMI_IR (eq. 2.1), and Downey *et al.*[11] did the same for the *Symmetric Conditional Probability* (SCP) (eq. 2.2).

$$PMI\_IR(a,b) = -\log \frac{\frac{hits(a\ AND\ b)}{total\_webs}}{\frac{hits(a)*hits(b)}{total\_webs}} \tag{2.1}$$

$$SCP\_IR(a,b) = \frac{\frac{hits(a\ AND\ b)}{total\_webs}}{\frac{hits(a)*hits(b)}{total\_webs}} \tag{2.2}$$

However, due to the lack of semantic background which may aid to properly interpret text, problems about language ambiguity (i.e. polysemic terms) or misinterpretation of co-occurrences (i.e. the type of semantic relation inherent to the term collocation cannot be evaluated) limit the performance of those approach with respect to ontology-based measures, as it will be shown in the evaluation section.

Precisely, ontology-based measures consider ontologies as an explicit domain knowledge graph model in which semantic interrelations are modeled as links between concepts. Basically, they exploit the taxonomic geometrical model(i.e. is-a links) to compute concept similarity, by measuring different features like concept inter-link distance (also called path length)[32] (eq. 2.3) and/or the depth of the taxonomy in which the concepts occur[17] (eq. 2.4).

$$sim_{path}(a,b) = min\ \#\ of\ is-a\ edges\ connecting\ a\ and\ b \tag{2.3}$$

$$sim_{leacock\&chodorow}(a,b) = -\log \frac{path(a,b)}{2 \times depth} \tag{2.4}$$

In the past, this idea has been applied to the MeSH semantic network, which consists of biomedical terms organized in a hierarchy[32]. Taking a similar approach, several authors[6,26] developed measures which compute path lengths in the UMLS hierarchy. The introduced measures have been also adapted by Ref.[28] to the biomedical domain by computing path lengths from the SNOMED-CT ontology. The advantage of this kind of measures is that they only use a domain ontology as the background knowledge, so, no corpus with domain data is needed. Their main problem is that they heavily depend on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology[8]. Moreover, it is worth to note that the presence of a semantic link between two concepts gives an evidence of a relationship but not about the strength of their semantic distance (i.e. all individual links have the same length and, in consequence, represent uniform distances[4]).

On the other hand, there exist other ontology-based similarity measures which combine the knowledge provided by an ontology and the Information Content (IC) of the concepts that are being compared. IC measures the amount of information

provided by a given term from its probability of appearance in a corpus. Consequently, infrequently appearing words are considered more informative than common ones.

Based on this premise, Resnik[29] presented a seminal work in which the similarity between two terms is estimated as the amount of information they share in common. In a taxonomy, this information is represented by the Least Common Subsumer (LCS) of both terms. So, the computation of the IC of the LCS results in an estimation of the similarity of the subsumed terms. The more specific the subsumer is (higher IC), the more similar the subsumed terms are, as they share more information. Several variations of this measure have been developed (as will be presented in section 3).

These measures have also been evaluated by Pedersen *et al.*[28] in the biomedical domain by using SNOMED-CT as ontology and a source of clinical data as corpus. As it will be shown in the evaluation section, they were able to outperform path length-based ones in this domain specific environment[28].

In the next sections, we present different approaches for IC-based similarity computation and we study how they can be modified in order to use the Web as a corpus, instead of specific clinical data (which may introduce applicability limitations as will be discussed in section 4).

3 IC-Based Similarity Measures

The *Information content* (IC) of a concept is the inverse to its probability of occurrence. The IC calculation is based on the probability *p(a)* of encountering a concept a in a given corpus, by applying eq 3.5. In this way, infrequent concepts obtain a higher IC than more common ones.

$$IC(a) = -\log p(a) \tag{3.5}$$

As mentioned above, Resnik[29] introduced the idea of computing the similarity between a pair of concepts (a and b) as the IC of their *Least Common Subsumer* (*LCS(a,b)*, i.e., the most concrete taxonomical ancestor common to a and b in a given ontology) (eq. 3.6). This gives an indication of the amount of information that concepts share in common. The more specific the subsummer is (higher IC), the more similar the terms are.

$$sim_{res}(a,b) = IC(LCS(a,b)) \tag{3.6}$$

The most widely used extensions to Resnik measure are Lin[20] and Jiang & Conrath[16].

Lin's[20] similarity depends on the relation between the information content of the LCS of two concepts and the sum of the information content of the individual concepts (eq. 3.7).

$$sim_{lin}(a,b) = \frac{2 \times sim_{res}(a,b)}{(IC(a) + IC(b))} \tag{3.7}$$

Jiang & Conrath[17] subtract the information content of the LCS from the sum of the information content of the individual concepts (eq. 3.8).

$$dis_{jcn}(a,b) = (IC(a) + IC(b)) - 2 \times sim_{res}(a,b) \tag{3.8}$$

Note that this is a dissimilarity measure because the more different the terms are, the higher the difference from their IC to the IC of their LCS will be.

Original works apply those measures by relying on WordNet[13] as the background ontology from where obtain the LCS of evaluated terms and SemCor[25] as a general purpose pre-tagged corpus from which obtain concept probabilities from manually computed term appearances. Thanks to the manual pre-processing of the corpus, concept probabilities are accurately computed and result in robust similarity estimations. However, in specific domains such as biomedicine, those general-purpose corpora present a reduced coverage[2], in addition to the limited coverage of biomedical terms of WordNet. These two issues result in a poor performance of the described similarity measures when applying them to concrete domain concepts[28]. Consequently, as stated in the previous section, Pedersen *et al.*[28] have been adapted them to the biomedical domain by exploiting SNOMED-CT taxonomy instead of WordNet and the Mayo Clinic Corpus of Clinical Notes corpus instead of SemCor.

On one hand, SNOMED-CT[1] (*Systematized Nomenclature of Medicine, Clinical Terms*) is an ontological/terminological resource publicly available distributed as part of the UMLS and it is used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerized physician order entry, disease surveillance, image indexing and consumer health information services. The SNOMED-CT ontology has a very good concept coverage[23,27,35] and it has been adopted as reference terminology by some countries, and some organizations[9]. It contains more than 311,000 medical concepts organized into 13 hierarchies with 1.36 million relationships, from which is-a relationships are exploited to extract the path and/or the LCS between a pair of terms.

On the other hand, the Mayo Clinic Corpus consists of 1,000,000 clinical notes collected over the year 2003 which cover a variety of major medical specialities at the Mayo Clinic. Clinical notes have a number of specific characteristics that are not found in other types of discourse, such as news articles or even scientific medical articles found in MEDLINE. Clinical notes are generated in the process of treating a patient at a clinic and contain the record of the patient-physician encounter. The notes were transcribed by trained personnel and structured according to the reasons, history, diagnostic, medications and other administrative information. Patient's history, diagnostic and medication notes were collected as the domain-specific and pre-processed data corpus from which to assess the semantic similarity between different pairs of diseases (more details in the evaluation section)[28].

4 Computing IC from the Web

Using a domain-specific and reliable corpus like the Mayo Clinic repository to compute IC-based semantic similarity may lead to very accurate results. However, the availability of those corpora (i.e. the use of patient data should ensure privacy and anonymity) and their coverage with respect to the evaluated terms (i.e. what happens if the evaluated concepts are not considered in typical clinical histories) are the main problems which hamper the applicability of those domain-dependant approaches. In fact, data sparseness (i.e. the fact that not enough data is available for certain concepts to reflect an appropriate semantic evidence) is the main problem of those approximations[5].

[1] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

On the other hand, the Web and, more particularly, web search engines are able to index almost any possible term. Considering its size and heterogeneity, it can be considered as a social-scale general purpose corpus. Its main advantages are its free and direct access and its wide coverage of any possible domain. In comparison with other general purpose repositories (such as the Brown Corpus or SemCor) which have shown a poor performance for domain-dependant problems[28], the Web's size is millions of orders of magnitude higher. In fact, the Web offers more than 1 trillion of accessible resources which are directly indexed by web search engines[2]. It has been demonstrated[5] the convenience of using such a wide corpus to improve the sample quality for statistical analysis.

In order to study the performance of the presented IC-based similarity measures with biomedical concepts when using the Web as a corpus, in this section, we adapt them by computing term occurrences from the Web instead of a reliable, closed and domain-specific repository of clinical data.

The main problem of computing term's Web occurrences is that the analysis of such an enormous repository for computing appearance frequencies is impracticable. However, the availability of massive Web Information Retrieval tools (general-purpose search engines like Google) can help in this purpose, because they provide the number of pages (hits) in which the searched terms occur. As introduced in section 2, this possibility was exploited in Ref.[36], by approximating PMI concept probabilities from web search engine hit counts.

This idea was later developed in Ref.[10], in which it is claimed that the probabilities of Web search engine terms, conceived as the frequencies of page counts returned by the search engine divided by the number of indexed pages, approximate the relative frequencies of those searched terms as actually used in society. So, exploiting Web Information Retrieval (IR) tools and concept's usage at a social scale as an indication of its generality, one can estimate, in an unsupervised fashion, the concept probabilities from Web hit counts.

Even though web-based statistical analyses brought benefits to domain-independent unsupervised approaches (i.e. no background ontology is exploited)[36], due to their lack of semantics, their performance is still far from the other supervised (ontology-based) measures[5] introduced in sections 2 and 3.

Taking those aspects into consideration, in the following, we will adapt the IC-based similarity measures introduced in section 3 to exploit the Web as a corpus, by estimating concept's IC from web hit counts. As stated, the LCS is extracted from the ontology. Following a similar principle as Turney[36], the Web-based IC computation is specified as follows.

$$IC\_IR(a) = -\log p_{web}(a) = -\log \frac{hits(a)}{total\_webs} \tag{4.9}$$

Being, $p_{web}(a)$ the probability of appearance of string 'a' in a web resource. This probability is estimated from the Web hit counts returned by Web IR tool -*hits*- when querying the term 'a'. *Total_webs* is the total number of resources indexed by a web search engine (estimated as 1 trillion, as stated above).

[2] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

In this manner, IC-based measures presented in section 3 can be directly rewritten to compute concept probabilities from the Web by incorporating the Web-based IC computation (IC_IR).

Resnik measure can be rewritten as follows.

$$sim_{res}\_IR(a,b) = IC\_IR(LCS(a,b)) = -\log \frac{hits(LCS(a,b))}{total\_webs} \tag{4.10}$$

Lin measure can be rewritten as follows.

$$\begin{aligned} sim_{lin}\_IR(a,b) &= \frac{2 \times sim_{res}\_IR(a,b)}{(IC\_IR(a) + IC\_IR(b))} = \\ &= \frac{2 \times (-\log \frac{hits(LCS(a,b)}{total\_webs})}{(-\log \frac{hits(a)}{total\_webs} - \log \frac{hits(b)}{total\_webs})} \end{aligned} \tag{4.11}$$

Finally, Jiang & Conrath distance measure can be rewritten as follows.

$$\begin{aligned} dis_{jcn}\_IR(a,b) &= (IC\_IR(a) + IC_I R(b)) - 2 \times sim_{res}\_IR(a,b) = \\ &= (-\log \frac{hits(a)}{total\_webs} - \log \frac{hits(b)}{total\_webs}) - 2 \times (-\log \frac{hits(LCS(a,b))}{total\_webs}) \end{aligned} \tag{4.12}$$

In any case, the main problem of using such a general repository as the Web to estimate concept's appearance probabilities is language ambiguity. This problem appears when concept probabilities are estimated by means of word's (instead of concept's) web hit counts. In comparison to unsupervised Web-based approaches, in this case, the introduction of the LCS extracted from the domain ontology helps to contextualize term occurrences towards the correct sense and forces and explicit taxonomic relation which leads to better similarity estimation.

However, even considering this advantage, on the one hand, different synonyms or lexicalizations of the same concept may result in different IC_IR values, introducing bias. On the other hand, the same word may have different senses and, in consequence, correspondences to several concepts. As a result, ambiguity may lead to inconsistent concept probability estimations. Fortunately, specific domain terms which are the object of this study, due to their concreteness, are rarely ambiguous in contrast to general words. So, our hypothesis is that language ambiguity will not compromise the web statistics when dealing biomedical concepts due to their high degree of concreteness. This is also evaluated in the next section.

5 Evaluation

The most common way to evaluate similarity measures is by using a set of word pairs whose similarity has been assessed by a group of human experts. Computing the correlation between the computerized and human-based ratings, one is able to obtain a quantitative value of the similarity function's quality, enabling an objective comparison against other measures. In a general setting, the most commonly used benchmark is the Miller and Charles set[24] of 30 ranked domain-independent word pairs, which is a subset of the benchmark, composed by 65 word pairs, proposed by Rubenstein and Goodenough[31].

For the biomedical domain, Pedersen *et al.*[28], in collaboration with Mayo Clinic experts, created a set of word pairs referring to medical disorders whose similarity

were evaluated by a group of 3 expert physicians. After a normalization process, a final set of 30 word pairs were selected with the corresponding averaged ratings provided by physicians in a scale from 1 to 4 (see Table 1). The correlation between human judgements was 0.68.

Table 1. Set of 30 medical term pairs with associated averaged expert's similarity scores (extracted from 28). Note that the term "*lung infiltrates*" is not found in SNOMED-CT.)

Term 1	Term 2	Physician score
Renal failure	Kidney failure	4.0
Heart	Myocardium	3.3
Stroke	Infarct	3.0
Abortion	Miscarriage	3.0
Delusion	Schizophrenia	3.0
Congestive heart failure	Pulmonary edema	3.0
Metastasis	Adenocarcinoma	2.7
Calcification	Stenosis	2.7
Diarrhea	Stomach cramps	2.3
Mitral stenosis	Atrial fibrillation	2.3
Chronic obstructive pulmonary disease	Lung infiltrates	2.3
Rheumatoid arthritis	Lupus	2.0
Brain tumor	Intracranial hemorrhage	2.0
Carpal tunnel syndrome	Osteoarthritis	2.0
Diabetes mellitus	Hypertension	2.0
Acne	Syringe	2.0
Antibiotic	Allergy	1.7
Cortisone	Total knee replacement	1.7
Pulmonary embolus	Myocardial infarction	1.7
Pulmonary fibrosis	Lung cancer	1.7
Cholangiocarcinoma	Colonoscopy	1.3
Lymphoid hyperplasia	Laryngeal cancer	1.3
Multiple sclerosis	Psychosis	1.0
Appendicitis	Osteoporosis	1.0
Rectal polyp	Aorta	1.0
Xerostomia	Alcoholic cirrhosis	1.0
Peptic ulcer disease	Myopia	1.0
Depression	Cellulitis	1.0
Varicose vein	Entire knee meniscus	1.0
Hyperlipidemia	Metastasis	1.0

Pedersen *et al.*[28] used that benchmark to evaluate several path-based and IC-based semantic similarity measures described in sections 2 and 3 respectively, exploiting the SNOMED-CT hierarchy (to compute paths and retrieve the LCS) and the Mayo Clinic Corpus (to statistically assess concept's IC). Note that the term pair "*chronic obstructive pulmonary disease*" - "*lung infiltrates*" was excluded from the test as the later term was not found in SNOMED-CT. The first 5 entries of Table 2 summarizes the correlations values they obtained in their tests for the different similarity measures.

In addition to Perdersen *et al.* experiments, we added the correlation values obtained for the two unsupervised Web-based measures introduced at the beginning of section 2 (PMI-IR and SCP-IR). Correlation values obtained for the same benchmark and using MS Bing[3] search engine are presented in the sixth and seventh entries of Table 2.

Finally, we compared these results for the same benchmark with those obtained by the three modified IC-based measures proposed in section 4, by substituting the domain-specific preprocessed corpus of Mayo Clinical Notes for the Web. We have also taken SNOMED-CT as the reference ontology from where obtain the required LCS. Again, in order to obtain term appearances from the Web we have used MS Bing as the web search engine. A priori, any other general-purpose search engine could be used (e.g. Google, Altavita or Yahoo). However, we have opted by Bing because it does not introduce limitations on the number of queries performed per day[34], and because we do not observed the hit count estimation inconsistencies of other search engines (like Google, more details in Ref.[33]), which may negatively affect the similarity assessment. The correlation results obtained for those tests are presented in the three last entries of Table 2.

Table 2. Correlations obtained for the different similarity measures against Perdersen's benchmark[28].

Measure	Knowledge and data sources	Correlation
Path	SNOMED-CT (Path)	0.36[28]
Leacock and Chorodow	SNOMED-CT (Path)	0.35[28]
Resnik	SNOMED-CT (LCS) and Mayo Clinical Notes (IC)	0.45[28]
Jiang and Conrath	SNOMED-CT (LCS) and Mayo Clinical Notes (IC)	0.45[28]
Lin	SNOMED-CT (LCS) and Mayo Clinical Notes (IC)	0.60[28]
PMI-IR	Web (collocation hit count)	-0.23
SCP-IR	Web (collocation hit count)	0.06
Resnik-IR	SNOMED-CT and Web (IC hit count)	0.48
Jiang and Conrath-IR	SNOMED-CT and Web (IC hit count)	0.59
Lin-IR	SNOMED-CT and Web (IC hit count)	0.63

It is worth to note that, due to the syntactical complexity of some of the LCSs extracted from SNOMED-CT (e.g. being "*morphologically altered structure*" the LCS of "*calcification*" and "*stenosis*") data sparseness may appear because a very few number of occurrences of the exact expression can be found ("*morphologically altered structure*" returned only 48 matches in Bing). In order to tackle this problem, we simplified syntactically complex LCSs by taking the noun phrase on the right of the expression or removing some adjectives on the left of the noun phrase (i.e. "*morphologically altered structure*" was replaced by "*altered structure*"). In this manner, as we generalize the LCS, the scope of the statistical assessment is increased (more hits are considered due to the simpler query expression) without losing much of the semantic context. Other approaches dealing with IC-based measures do not face this problem, as term frequencies of the LCS are manually computed from the background corpus at a conceptual level[16] (i.e. a document may cover the "*morphologically altered structure*" topic even though the exact term expression never appears in the

[3] http://www.bing.com [Accessed on January 12th, 2010]

text, a situation which is detected and taken into consideration by a human expert). This contrasts to the strict keyword-matching of standard Web search engines which only seek for the presence or absence of the query expression into the text.

Analysing the results presented in Table 2 we can draw several conclusions. Regarding the experiments performed by Pedersen *et al.* one can see that path-based measures are easily surpassed by IC-based measures as the latter exploit more semantic evidences that the former. In that case, the use of a repository from which retrieve concept appearance frequencies result in more reliable results in comparison to human judgements. All values are, as expected, below the 0.68 correlation obtained between the human judgements, which represents an upper bound for a computerized approach.

In comparison, the two Web-based unsupervised measures (PMI-IR and SCP-IR) performed very poorly, with correlations near to 0 or even lower (i.e. worse than random assessments). It is important to note that those functions do not use any ontology as background and they uniquely rely on the degree of term co-occurrence obtained from web search engines. Moreover, due to their lack of semantic background which result in uncontextual web queries, they are more aimed to compute word relatedness rather than concept similarity as they are not able to assess the kind of semantic relationship inherent to term co-occurrence. The main problem here is that terms are so concrete that their explicit co-occurrence in a web resource is very rare (even though being certainly related). As a result, many queries result in zero hits, underestimating the semantic relationship of both concepts. The assessment problems related on uniquely relying on explicit term co-occurrences where also stated in Ref.[19]. So, even though those unsupervised functions performed reasonably well in a general setting not affected by data sparseness (like in the evaluations performed by Ref.[36] and Ref.[11]) they fail to distinguish the subtle nuances of very concrete domain terms like those composing Pedersen *et al.* benchmark.

From the above results, one may conclude that term hit count obtained from the Web are not reliable enough to estimate concept's distribution in an accurate way. However, observing the results obtained by the three last IC-based measures, which employ those hit counts to estimate concept appearance frequencies, the conclusion is the opposite. In this case, similarity values computed from the Web correlate better than even those obtained from a domain-specific pre-processed corpus. In some cases (Jiang & Conrath measure), the improvement is around a 20% (0.45 vs 0.59) with respect to the upper bound and, in others (Lin measure), results are very close to the human judgements (with a correlation of 0.63 vs 0.68).

That is a very interesting conclusion, showing that, one the one hand, the exploitation of a domain ontology leads to appropriate term queries which can be appropriately evaluated in the Web (i.e. as concepts are evaluated individually, there are not data sparseness problems derived from the requirement of an explicit co-occurrence); on the other hand, the fact of exploiting the LCS as an indication of the concept's taxonomic relationship (i.e. similarity) leads to much better results than unsupervised measures introduced above. Finally, the inclusion of the LCS in the query allows focusing the retrieval towards more appropriate resources than when using an uncontextual query. This constraints the Web analysis towards those resources which are related to the domain (in this case, medical ones) rather than the whole

Web, increasing the effectiveness of the assessment.

In this manner, the Web (accessed by means of publicly available web search engines), despite its generality, noise, lack of structure and unreliability of individual sources is able to provide a robust semantic evidence for concrete domains such as biomedicine. In fact, the Web provided a better assessment of similarity than when using relevant (but reduced) domain corpora. As stated in the introduction, the size and heterogeneity of the Web aids to provide accurate estimations of domain information distribution at a social scale, which improves the distribution observed in a much more reliable and structured, but also reduced and potentially biased, source.

In addition, even though polysemy and synonymy may affect the computation of concept probabilities from word web hit counts (due to the limitations of the strict keyword matching algorithms implemented by search engines, as stated in sections 4 and 5), thanks to the reduced ambiguity of concrete domain-specific words, results do not seem to be seriously affected.

6 Conclusions

In this paper, we presented several semantic similarity computation paradigms and evaluated the influence of the background corpus used by ontology-driven IC-based similarity measures when applied to a specific domain such as biomedicine.

In previous approaches, it was argued that, for supervised ontology-based similarity measures, a domain-specific corpus was needed to achieve reliable similarity values for domain-specific concepts[28]. Considering the characteristics of the Web and its success in previous attempts of exploiting it to tackle other language processing tasks[36], we adapted IC-based measures to compute concept probabilities from term web search hit counts.

Coherently to the hypothesis stated in this paper, the evaluation has shown that using the Web, despite of being a priori less reliable, noisier and unstructured, similarity values are even more reliable (compared to human judgments) than those obtained from a domain-specific corpus (maintaining the same domain ontology, SNOMED-CT, as background knowledge). Moreover, the limitations of the strict keyword matching implemented by web search engines have not handicapped the results, providing better estimations than appearance frequencies computed from pre-processed domain data. Consequently, in this case, the necessity of having a domain corpus is no more required. This is a very interesting conclusion because, usually, domain corpus lacks of coverage, it has a reduced size or even it is not available due to the confidentiality of data. This was observed however, only for supervised measures, as unsupervised approaches were hampered by the data sparseness caused by the necessity of explicit co-occurrence of concrete domain terms and the ambiguity of those co-occurrences. This was expectable, as previous works have also stated the limitations of unsupervised similarity measures in comparison to supervised ones[15]. In this case, moreover, the concreteness of the benchmark amplified the differences even more.

Summarizing, from the experiments, we can conclude that the Web (in a raw and unprocessed manner) is a valid corpus from which to compute ontology-based semantic similarities in a concrete domain as biomedicine.

After this work, we plan to evaluate in the Web the proposed IC-based similarity measures in other concrete domains for which ontologies are available (such as

chemistry or computer science). We also studied how they performed in a general environment with domain independent -and potentially ambiguous- terms[33]. In that last case, it was observed that the result's quality was compromised by the inaccurate estimation of web-based concept probabilities caused by language ambiguity. So, additional strategies were needed in order to contextualize queries for concept probability estimation by exploiting available ontological knowledge (i.e. attaching the LCS to every web query)[33].

References

[1] Aseervatham S, Bennani Y. Semi-Structured document categorization with a semantic kernel. Pattern Recognition, 2009, 42(9): 2067–2076. [doi: 10.1016/j.patcog.2008.10.024]

[2] Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. Proceedings of the NAACL (North American Association for Computational Linguistics) 2001 Workshop: WordNet and other lexical resources: Applications, extensions and customizations, Pittsburgh, PA, 2001. 77–82.

[3] Budanitsky A, Hirst G. Evaluating wordnet-based measures of semantic distance. Computational Linguistics, 2006, 32(1): 13-47.

[4] Bollegala D, Matsuo Y, Ishizuka M. WebSim: A Web-based Semantic Similarity Measure. The 21st Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2007). Miyazaki, Japan, 2007. 757–766.

[5] Brill E. Processing Natural Language without Natural Language Processing. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico, 2003. 360–369.

[6] Caviedes J, Cimino J. Towards the development of a conceptual distance metric for the UMLS. Journal of Biomedical Informatics, 2004, 37: 77–85. [doi: 10.1016/j.jbi.2004.02.001]

[7] Church KW, Gale W, Hanks P, *et al.* Using Statistics in Lexical Analysis. Proc. of Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, 1991. 115–164.

[8] Cimianio P. Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. Springer-Verlag, 2006.

[9] Cornet R, Keizer NF. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making, 2008, 8(Suppl 1): S2. [doi: 10.1186/1472-6947-8-S1-S2]

[10] Cilibrasi RL, Vitáyi PMB. The Google Similarity Distance. IEEE Transaction on Knowledge and Data Engineering, 2006, 19(30): 370–383.

[11] Downey D, Broadhead M, Etzioni O. Locating complex named entities in Web text. In: Proc. of the 20th International Joint Conference on Artificial Intelligence, 2007. 2733–2739.

[12] Etzioni O, Cafarella M, Downey D, *et al.* Unsupervised named-entity extraction form the Web: An experimental study. Artificial Intelligence, 2005, 165: 91–134.

[13] Fellbaum C. WordNet: An Electronic Lexical Database, MIT Press. More information: http:// www.cogsci.princeton.edu/ wn/, Cambridge, Massachusetts, 1998.

[14] Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EGM, Milios EE. Information Retrieval by Semantic Similarity. International Journal on Semantic Web and Information Systems, 2006, 2(3): 55–73.

[15] Iosif E, Potamianos A. Unsupervised Semantic Similarity Computation using Web Search Engines. In: Proc. of the International Conference on Web Intelligence, 2007. 381–387.

[16] Jiang J, Conrath D. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Proc. of the International Conference on Research in Computational Linguistics (ROCLING X), Taiwan, Sep 1997. 19–33.

[17] Leacock C, Martin C. Combining local context and WordNet similarity for word sense identification. chapter WordNet: An electronic lexical database. MIT Press, 1998. 265–283.

[18] Landauer TK, Dumais ST. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review, 1997, 104: 211–240.

[19] Lemaire B, Denhière G. Effects of High-Order Co-occurrences on Word Semantic Similarities. Current Psychology Letters - Behaviour, Brain and Cognition, 18(1): http://cpl.revues.org/document471.html

[20] Lin D. An Information-Theoretic Definition of Similarity. Proc. of the 15th International Conference on Machine Learning (ICML98). Madison, Wisconsin, USA, 1998, Morgan Kaufmann. 296–304.

[21] Lee, JH, Kim MH, Lee YJ. Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation, 1993, 49(2): 188–207.

[22] Lord P, Stevens R, Brass A, Goble C. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics, 2003, 19(10): 1275–1283. [doi: 10.1093/bioinformatics/btg153]

[23] Lieberman, MI, Ricciardi TN, Masarie FE, Spackman KA. The use of SNOMED CT simplifies querying of a clinical data warehouse. AMIA Annual Symposium Proceedings, 2003. 910.

[24] Miller GA, Charles WG. Contextual correlates of semantic similarity. Language and Cognitive Processes, 1991, 6(1): 1–28.

[25] Miller G, Leacock C, Tengi R, *et al.* A Semantic Concordance. HLT'93: Proceedings of the workshop on Human Language Technology, 1993. Association for Computational Linguistics. 303–308.

[26] Liu JZ, Wang W, Yang J. A framework for ontology-driven subspace clustering. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. 623–628.

[27] Penz JF, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, Lincoln MJ. Evaluation of SNOMED coverage of Veterans Health Administration terms. In: 11th World Congress on Medical Informatics, Medinfo. IOS Press, 2004. 540–544.

[28] Pedersen T, Pakhomov, S, Patwardhan S, Chute C. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics, 2007, 40: 288–299. [doi: 10.1016/j.jbi.2006.06.004]

[29] Resnik P. Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95), Montreal, Canada, 1995. 448–453.

[30] Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 1999, 11: 95–130.

[31] Rubenstein H, Goodenough J. Contextual correlates of synonymy. Communications of the ACM, 1965, 8(10): 627–633.

[32] Rada R, Mili H, Bichnell E, Blettner M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 1989, 9(1): 17–30.

[33] Sánchez D, Batet M, Valls A, Gibert K. Ontology-driven web-based semantic similarity. Journal of Intelligent Information Systems, 2009. (In press) [doi: 10.1007/s10844-009-0103-x]

[34] Sánchez D, Moreno A. Learning non-taxonomic relationships from web documents for domain ontology construction. Data Knowledge Engineering. Elsevier, 2008, 63(3): 600–623. [doi: 10.1016/j.datak.2007.10.001]

[35] Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthcare Informatics, 2004, 21(9): 54–56.

[36] Turney PD. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proc. of the Twelfth European Conference on Machine Learning, Freiburg. Germany, 2001. 491–499.

[37] Wilbu W, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Computers in Biology and Medicine, 1996, 26: 209–222.

[38] Wu ZB, Palmer M. Verb semantics and lexical selection. In: Proc. of the 32nd annual Meeting of the Association for Computational Linguistics, New Mexico, USA, 1994. Association for Computational Linguistics. 133–138.