Determining the Degree of Semantic Similarity Using Prototype Vectors*

Mireya Tovar[1,2], David Pinto[2], Azucena Montes[1,3], and Darnes Vilariño[2]

[1] Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Mexico
[2] Faculty Computer Science, Benemérita Universidad Autónoma de Puebla, Mexico
[3] Engineering Institute, Universidad Nacional Autónoma de Mexico
{mtovar,amontes}@cenidet.edu.mx,
{dpinto,darnes}@cs.buap.mx

Abstract. Measuring the degree of semantic similarity for word pairs is very challenging task that has been addressed by the computational linguistics community in the recent years. In this paper, we propose a method for evaluating input word pairs in order to measure the degree of semantic similarity. This unsupervised method uses a prototype vector calculated on the basis of word pair representative vectors which are contructed by using snippets automatically gathered from the world wide web.

The obtained results shown that the approach based on prototype vectors outperforms the results reported in the literature for a particular semantic similarity class.

Keywords: Semantic similarity, hierarchical relationships, prototype vectors.

1 Introduction

With the exponential growth of the information contained in the World Wide Web it arises the need for automating user processes such as searching, information retrieval, question answering, etc. One of the main problems of this automation is that much of the information remains unstructured, i.e., it is written in natural language and its ambiguity is difficult to be automatically processed by machines. The Semantic Web attempts to solve these problems by incorporating semantic to the web data, so that it can be processed directly or indirectly by machines in order to transform it into a data network [1]. For this purpose, it has been proposed to use some knowledge structures such as ontologies for giving semantic and some structure to unstructured data. Among other applications, an ontology is a lexical/semantic resource that allows to perform semantic annotation of web pages contents. Thus, we consider very important to investigate the manner of evaluate the quality of these kind of resources that are continuously been used in the framework of semantic web.

* This work is partialy supported by CONACYT and PROMEP under grants: CONACYT 54371, PROMEP/103.5/12/4962 BUAP-792 and project CONACYT 106625.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2013, LNCS 7914, pp. 364–373, 2013.

Gruber [2] defines an ontology as: "an explicit specification of a conceptualiation". An ontology includes classes, instances, attributes, relationships, constraints, rules, events and axioms. In many cases, ontologies are structured as hierarchies of concepts modeled either by means of part-whole or class-inclusion semantic relationships. In particular, the class-inclusion relationships are also named is-a, hyponymy or subsumption. There exist, others type of semantic relationships that are not hierarchical such as synonyms, antonyms, etc, however, in this paper, we focus the experiments in the evaluation of semantic hierarchical relationships.

There are two types of nodes in an ontology: concepts and instances [3], but in this paper, we are particularly interested in determining the degree of similarity between a given pair of instances of the ontology that share a semantic hierarchical relationship.

A number of diverse classification methods have been addressed for identifying relationships between concepts or instances [4], [5] and [6]. For instance, for identifying whether or not a given instance (a pair of words *flower*:*tulip*) belongs to a specific relationship (class-inclusion) [7].

Other approaches identify the degree of semantic similarity between a set of word pairs that is known that they belong to a certain class (semantic relationship) [8].

The latter case is the matter of this research work and clearly this problem goes beyond of identifying the membership of an instance in a given class, that is, to detect the variability of the instance with respect to the class.

The remaining of this paper is structured as follows. Section 2 describes with more detail the problem of measuring the degree of semantic similarity for hierarchical relationships. The state of the art is also discussed in Section 2. In Section 3, we present the model proposed for addressing the problem aforementioned. Section 4 show and discuss the results obtained by the presented approach. Finally, in Section 5 the findings and the future work are given.

2 Degree of Semantic Similarity

Even if there exist a number of widely used semantic relationships (see Table 1), in this paper we consider only two classes: Class-Inclusion and Part-Whole for determining the degree of semantic similarity between two instances of an ontology.

The degree of semantic similarity involves the process of determining a ranking for pairs that belong to the same semantic class. For instance, let us consider the following word pairs: $\{dog : bark\}$, $\{cat : meow\}$ and $\{floor : squeak\}$ that share the ENTITY:SOUND semantic relationship. The intuition is that the first pair is more similar to the second than with the third one. In this sense, it is very important to construct formulae that allows to rank word pairs sharing the same semantic similarity. In this paper, we analize techniques for obtaining such ranking by means of prototype vectors. For the experiments carried out, we have considered only two semantic relationships: class-inclusion and part-whole. A description of these two relationships follows.

Table 1. High-level classes and instances examples that belong to some subclass [9]

	High-level class	Subclass	Instances examples
1	**CLASS-INCLUSION**	Taxonomic	flower:tulip
2	**PART-WHOLE**	Object:Component	car:engine
3	SIMILAR	Synonymity	car:auto
4	CONTRAST	Contradictory	alive:dead
5	ATTRIBUTE	Item:Attribute	beggar:poor
6	NON-ATTRIBUTE	Item:Nonattribute	harmony:discordant
7	CASE RELATIONS	Agent:Recipient	doctor:patient
8	CAUSE-PURPOSE	Cause:Effect	joke:laughter
9	SPACE-TIME	Location:Action/Activity	school:learn
10	REFERENCE	Sign:Significant	siren:danger

The Class-Inclusion relationship defines a parent-child relationship (taxonomy, e.g. flower: tulip), whereas part-whole relationship divides a concept as a whole in different parts (e.g engine:car). Both relationships, as mentioned before, are considered to be hierarchical [10], [11], [12]. The subclasses of the Class-Inclusion and the Part-whole classes are shown in Table 2 and Table 3, respectively. Syntactic patterns for each subclass, together with instance examples are also presented in these Tables.

Table 2. The **Class-Inclusion** class: subclasses, syntactical patterns and examples of instance pairs that belong to each subclass

	Subclass	**Syntactical patterns**	**Examples of instances ($Y : X$)**
1b	Functional	Y functions as an X	ornament:brooch, weapon:knife, vehicle:car
1c	Singular Collective	a Y is one item in a collection/group of X	cutlery:spoon, clothing:shirt, vermin:rat, medicine:asprin
1d	Plural Collective	Y are items in a collection/group of X	groceries:eggs, refreshments:sandwiches, drugs:amphetamines
1e	ClassIndividual	Y is a specific X	queen:Elizabeth, river:Nile, city:Berlin

One of the most important forums that have tackled the problem of identifying the degree of semantic similarity is Semeval[1]. There have been some teams that presented different approaches for this particular problem. Thus, the state of the art is scarce but it follows.

In [13] it was proposed two systems that tackled the problem. Their methodology employed lexical patterns generated from the contexts in which the word pairs occurs. They constructed patterns using the example word pairs for each subclass (see the third column of Table 2). They used a corpus with 8.4 million

[1] `http://www.cs.york.ac.uk/semeval-2012/`

Table 3. The **Part-Whole** class: subclasses, syntactical patterns and examples of instance pairs that belong to each subclass

Subclass	**Syntactical patterns**	**Examples of instances** ($X : Y$)
Object:Component	a Y is a part of an X	car:engine, face:nose, novel:epilogue, turtle:carapace
Collection:Member	X is made from a collection of Y	forest:tree, anthology:poem, fleet:ship, medley:melodies
Event:Feature	Y is typically found at an event such as X	rodeo:cowboy, banquet:food, wedding:bride
Activity:Stage	X is one step/action/part of the actions in Y	shopping:buying, planting:gardening, kickoff:football
Item:Topological Part	Y is one of the areas/locations of X	room:corner, mountain:foot, table:top
Object:Stuff	X is made of / is comprised of Y	glacier:ice, salt:sodium, lens:glass, parquet:wood
Item:Distinctive Nonpart	X is devoid of / cannot have Y	tundra:tree, horse:wings, perfection:fault, society:pariah
Item:Ex-part/Ex-possession	an X once had/owned/possessed Y but no longer	apostate:belief, wood:splinter, prisoner:freedom, metal:dross

documents from Gigaword and 4 million articles from Wikipedia. They ranked word pairs using a model predicting the probability that they belong to the input relationship. They proposed two approaches for ranking the word pairs of the subclasses: The **UTD-NB** approach used a probabilistic model, whereas **UTD-SVM** employed a SVM-rank model to rank the word pairs. Their performance was interesting, achieving good results reported in the Semeval conference [8].

In [14] it was proposed three unsupervised approaches that used the Gloss Vector measure found in the package WordNet::Similarity. The author expanded the vector of glosses by using the relationships associated to each word pair. Additionally, superglosses have been produced. The cosine measure was employed for ranking the results obtained for each pair of words. The corpus used was the complete collection of glosses and examples from WordNet 3.0, i. e., 118,000 glosses.

In [15] a supervised approach has been proposed. They used lexical, semantic, WordNet-based and contextual features. In order to rank the obtained results, the cosine measure was employed. Moreover, they used a very restricted corpus for testing their approach in comparison with the other teams of the competition.

The approach presented in this paper is described in the following section.

3 A Prototype-Based Model

In order to calculate a ranking for a set of word pairs that belong to the same semantic relationship, we have constructed a prototype vector for this relationship. This prototype is calculated as the average value among all the representative vectors for each word pair. We assume that the prototype vector (also known as the class centroid) has enough quality for representing such semantic relationship. Thus, the final ranking is calculated by means of the distance that exist between each word pair representative vector and the prototype vector.

The representative vectors for each word pair are calculated in the basis of information gathered from the Web by using a search engine[2]. Thus, we obtain snippets that contain the two words of the given pair. All the snippets for each word pair are analized for constructing a representative vector using the TF-IDF weighting schema. In Figure 1 we show the unsupervised approximation for determining the degree of instance prototypicality within a given subclass. As we mentioned before, we only considered two classes (class-inclusion and part-whole), and 12 ($n = 12$) subclasses, as shown in Tables 2 and 3.

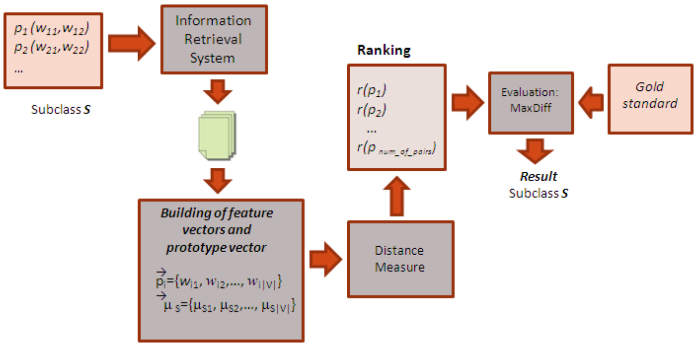

Fig. 1. The proposed architecture for calculating the degree of semantic similarity

The architecture proposed presents the following phases:

- *Gathering the corpus.* The corpus consists of short texts (snippets) containing the keywords of the instances (word pairs). We used the Google API for gathering such snippets from the world wide web.
- *Building the feature vectors.* Using the vocabulary of the corpus, we built the feature vectors for each subclass. Normally, each subclass contain between 27 to 31 instance pairs, and the feature vector contains frecuency-based lexical features of the terms associated the instance pairs. Let V the vocabulary of the whole corpus, and $\vec{p_i}$ the representative vector for the word pair p_i. Thus, $\vec{p_i} = \{w_{i,1}, w_{i,2}, \cdots, w_{i,|V|}\}$, where $w_{i,k}$ is the weight of the k-th element of the vocabulary in the $\vec{p_i}$ representative vector of the word pair p_i.

[2] In this case we used Google.com

This weight is calculated using two formulae: term frequencies (TF) and the combination of term frequency and inverse document frequency (TF-IDF).

– *Constructing the prototype vector.* The prototype vector is constructed by considering the mean value among all the feature vectors for all the possible word pairs in a given subclass. Thus, this prototype vector may be seen as the centroid of the subclass. Formally, for each $p_i \in S$ with $i = \{1, \cdots, num\_of\_pairs\}$, where S is the current subclass, the centroid μ_S of this subclass is calculated as shown in Eq.1.

$$\overrightarrow{\mu_S} = \{\mu_{S,1}, \mu_{S,2}, \cdots, \mu_{S,|V|}\}$$
$$\text{with } \mu_{S,k} = \frac{1}{num\_of\_pairs} \sum_{i=1}^{num\_of\_pairs} w_{i,k} \tag{1}$$

– *Calculating the ranking* We apply different distance measures between the feature vectors and their corresponding prototype vector for determining a score that indicates the degree of representativeness of the instance pairs with respect to the whole subclass. Thus, the closer a representive vector to the centroid, the higher the ranking of representativeness. In other words, we assume that those word pairs that are closer to the subclass prototype vector are more representative for this subclass.

  In order to measure the distance among the representative vectors and their corresponding prototype vector, we employed two classical distance measures: Euclidean (Equation 2) and Manhattan (Equation 3).

$$d(\overrightarrow{p_i}, \overrightarrow{\mu_S}) = \frac{\|\overrightarrow{p_i} - \overrightarrow{\mu_S}\|}{|V|} \tag{2}$$

$$d(\overrightarrow{p_i}, \overrightarrow{\mu_S}) = \frac{\sum_{k=1}^{|V|} |p_{i,k} - \mu_{S,k}|}{|V|} \tag{3}$$

– *Evaluating the ranking.* In order to determine the performance of the presented approach, we have used the MaxDiff evaluation measure. MaxDiff is an analitycal techique that indicate the preference that a respondent has for a set of alternatives [16], [17]. This measure provides an average evaluation result, when evaluating the score of each pair of instances of the subclass with respect to a given gold standard. This reference of evaluation (gold standard) is built as follows. The respondents are asked to evaluate four or five pairs of words in an specified subclass and later they choose the best and worst pair of instance prototipical of the relationship [8].

4 Experimental Results

In this section we present the results obtained with the proposed approach. These results are compared with those reported in [8]. We have reported only the 12 subclasses that we have analized in this paper. We first describe the dataset used in the experiments, and later the obtained results are presented and discussed.

4.1 Dataset

Tables 4 and 5 present the number of word pairs to be evaluated for each subclass, the number of possible swaps (reversals), and examples of word pairs. The reversals were added to the subclass with the aim of determining how efficient are the approaches that consider these reversals pairs. In particular, with relationships that are completely directional such as Part-Whole, but also for other type of simetric classes, such as synonymy.

An example of a reversal word pair could be *dogs:animals* and *animals:dogs* of the Part-Whole class and the Plural Collective subclass. The ranking for each one should be different because the relationship is valid in only one direction. In this example, the correct word pair should be *animals:dogs*, because the representative pattern "Y are items in a collection/group of X" fulfills the relationship semantics when the pattern is used.

Table 4. The **Class-Inclution** class: subclasses, number of pairs and examples of instance pairs that belong to each subclass to evaluate

Subclass	Number of pairs	Reversals	Example
Functional	41	6	instrument:clarinet, fuel:gasoline, seat:stool, seat:chair, lubricant:oil, home:tree
Singular Collective	42	5	internet:website, book:novel, beverage:water,fruit:apple, art:sculpture
Plural Collective	43	5	birds:crows, colors:blue, silverware:spoon, dogs:animals animals dogs
ClassIndividua	33	2	horse:Palomino, snake:Cobra, Earth:planet, king:Arthur, university:Yale

4.2 Results

We gathered approximately 264,000 snippets from Internet in order to made up the reference corpus for the 12 subclasses to be evaluated. Table 6 and 7 show the average performance obtained by the four different runs executed in the experiments using the Class-Inclusion and the Part-Whole classes, respectively. The other results correspond to those reported in the literature. In particular, the run named **Euclidian-TF** uses **TF** as a weighting schema, employing the Euclidian distance measure for ranking the degree of semantic similarity. **Euclidian-TF-IDF** uses **TF-IDF** as a weighting schema in combination with the Euclidian distance measure. The runs named **Manhattan-TF** and **Manhattan-TF-IDF** are similar but they use the Manhattan distance measure instead of the Euclidian one.

From the obtained results we can observe that there is not a significative difference between the two distance measures employed. Actually, it can be seen that the Euclidean distance generally improves the Manhattan one. This means that the snippets that correspond to a given word pair that were gathered from

Table 5. The **Part-Whole** class: subclasses, number of pairs and examples of instance pairs that belong to each subclass to evaluate

Subclass	Number of pairs	Reversals	Example
Object:Component	44	5	house:room, recipe:ingredient, fin:fish, motor:boat, eye:lashes
Collection:Member	38	4	flock:sheep, album:songs, paragraph:word, herd:antelopes
Event:Feature	39	3	funeral:coffin, church:preacher, baptism:priest, competition:athlete
Activity:Stage	40	5	soaping:showering, stitching:sewing, purling:knitting, tennis:volleying
Item:Topological Part	43	4	tree:root, river:bed, coast:east, bush:roots
Object:Stuff	43	5	boots:leather, lawn:grass, box:cardboard, sock:thread
Item:Distinctive Nonpart	39	4	forest:sand, hearing:deaf, pride:embarrassment, venus:life
Item:Ex-part/Ex-possession	42	4	corpse:life, repair:break, notebook:paper, widow:husband

Table 6. Results for CLASS-INCLUSION Class

Approaches	MaxDiff
Euclidian-TF	**39.15**
Manhattan-TF	**38.85**
UTD-NB	37.60
UMD-V1	35.60
UMD-V2	33.13
UTD-SVM	31.58
Euclidian-TF-IDF	**31.55**
BUAP	31.43
Random	30.98
Manhattan-TF-IDF	**30.60**
UMD-V0	29.23

the Internet are quite similar because the representative vectors are close to the prototype vector (it is a dense group of representative vectors). This behaviour is valid for approximately 39% of the word pairs (for the Class-Inclusion class), whereas this percentage was approximately 30% for the Part-Whole class. The rest of word pairs need to be tuned in order to retrieve more representative snippets, or a new document collection which may be used for the same purpose.

It is remarkable that the **TF** representation schema obtained a very good performance with the Class-Inclusion class. We consider that this behavior is due to the unsupervised system proposed do not consider symmetric vs asymmetric relationships (reversal word pairs). The proposed system neither consider

Table 7. Results for the PART-WHOLE Class

Approaches	MaxDiff
UTD-NB	40.89
UTD-SVM	35.65
BUAP	35.05
Euclidian-TF-IDF	**33.30**
Manhattan-TF-IDF	**32.13**
Random	31.86
Manhattan-TF	**30.45**
Euclidian-TF	**29.46**
UMD-V0	29.40
UMD-V2	28.55
UMD-V1	26.51

the order of the words in the relationship, nor the direction of the semantic relationship. However, both Class-Inclusion and Part-Whole are asymmetric relationships and the approach presented is highly sensible to reversals, therefore, the performance decreases. This consideration is also avoided in the information retrieval system leading to obtain snippets that may not represent adequately the semantic relationship. Further investigation should integrate the concept of symmetric vs asymmetric relationships into the methodology proposed.

In summary, the results obtained outperformed all the results reported in the literature for the Class-Inclusion class, whereas in the class Part-Whole, we have slightly improved the random baseline. Further experiments will allow to analize the manner we may improve these results.

5 Conclusions

The schema proposed for measuring the degree of similarity based on prototype vectors performed well for the Class-Inclusion class. A simple weighting measure such as term frequencies allowed to capture the neccesary features for representing the word pairs that share a semantic relationship. On the other hand, we did not succeed representing adequately for the Part-Whole class, an issued that must be further investigated.

From the results obtained, we were able to observe that the snippets retrieved have an acceptable quality, but they may also be improved by adding other lexical resources or document collections. As future work, we would like to analize more into detail the outliers in order to remove them for generating a much better reference corpus.

References

1. Solís, S.: La Web Semántica. Lulu Enterprises Incorporated (2007)
2. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL-93-04, Knowledge Systems Laboratory, USA (1993)

3. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 67–79. Springer, Heidelberg (2005)
4. Celli, F.: Unitn: Part-of-speech counting in relation extraction. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 2010, pp. 198–201. Association for Computational Linguistics, Stroudsburg (2010)
5. Chen, Y., Lan, M., Su, J., Zhou, Z.M., Xu, Y.: Ecnu: Effective semantic relations classification without complicated features or multiple external corpora. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 2010, pp. 226–229. Association for Computational Linguistics, Stroudsburg (2010)
6. Negri, M., Kouylekov, M.: Fbk_nk: A wordnetbased system for multiway classification of semantic relations. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 2010, pp. 202–205. Association for Computational Linguistics, Stroudsburg (2010)
7. Turney, P.D.: Similarity of semantic relations. Computational Linguistics 32, 379–416 (2006)
8. Jurgens, D.A., Mohammad, S.M., Turney, P.D., Holyoak, K.J.: Semeval-2012 task 2: Measuring degrees of relational similarity. In: Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval 2012, Montreal, Canada (2012)
9. Bejar, I., Chaffin, R., Embretson, S.: Cognitive and psychometric analysis of analogical problem solving. Recent Research in Psychology. Springer (1991)
10. Lim, E.: Intelligent Systems Reference Library: Knowledge Seeker: Ontology Modelling for Information Search and Management - A Compendium. Intelligent Systems Reference Library. Springer (2011)
11. Harpring, P.: Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. Introduction To Series. Getty Research Institute (2010)
12. Organization, N. (U.S.), N.I.S.O.: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. National information standards series. NISO Press (2005)
13. Rink, B., Harabagiu, S.: Utd: Determining relational similarity using lexical patterns. In: Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval 2012, June 7-8, pp. 413–418. Association for Computational Linguistics, Montréal (2012)
14. Pedersen, T.: Duluth: Measuring degrees of relational similarity with the gloss vector measure of semantic relatedness. In: Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval 2012, June 7-8, pp. 497–501. Association for Computational Linguistics, Montréal (2012)
15. Tovar, M., Reyes, J.A., Montes, A., Vilariño, D., Pinto, D., León, S.: Buap: A first approximation to relational similarity measuring. In: Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval 2012, June 7-8, pp. 502–505. Association for Computational Linguistics, Montréal (2012)
16. Karash, K., Board, N.R.C.U.T.R., Program, T.C.R., Administration, U.S.F.T., Corporation, T.D.: Understanding How Individuals Make Travel and Location Decisions: Implications for Public Transportation. Report (Transit Cooperative Research Program). Transportation Research Board (2008)
17. Orme, B.: Maxdiff analysis: Simple counting, individual-level logit, and hb. Technical report, Sawtooth Software. Research Paper Series (2009)