Quantifying Relational Triples

Hernani Costa[1] and Hugo Gonçalo Oliveira[1⋆] Paulo Gomes[1]
hpcosta@student.dei.uc.pt, hroliv@dei.uc.pt, pgomes@dei.uc.pt

Cognitive and Media Systems Group (CISUC)
University of Coimbra, Portugal

Abstract. The evaluation of knowledge is a very challenging task, which generally ends up being done by humans. Despite less prone to errors, manual evaluation is hardly repeatable, time-consuming and sometimes subjective. In this paper, we propose to quantify relational triples automatically, exploiting popular distributional similarity measures. In the first experiment, we used these measures to quantify triples according to the co-occurrence of their arguments in text. In the second, we attached textual patterns denoting their relation and used the Web to validate them. In both experiments some scores revealed to be highly correlated with the quality of the triples.

1 Introduction

Tasks like information extraction (IE) or information retrieval (IR), where is necessary to understand the interactions between the words of a language and its meaning, lead to the creation of broad-coverage computational semantic resources, like lexical ontologies (e.g. Wordnet [10]). However, the creation and maintenance of this kind of resource involves too much human work. Therefore, during the last decades, there have been several attempts to discover knowledge automatically from text (see for instance [16] [4] [21] [6]). These approaches have been applied to different types of text, and the lexical-semantic knowledge has been extracted from structured resources, like dictionaries [13], or non-structured resources, like corpus [4]. Regardless the kind of knowledge, IE systems generally acquire entities (e.g. e_1, e_2) and relations between them, represented as triples (e_1, r, e_2), where r identifies the type of relation.

Knowledge discovered automatically is useful to create or to enrich existing ontologies, however its evaluation is quite a challenging task, especially when dealing with broad-coverage open-domain knowledge. Even though time-consuming, hardly repeatable and sometimes subjective, manual evaluation of a representative set of knowledge is the most common choice (as in [4] [21]).

Nevertheless, an alternative is to search, in large collections of text, for support on the knowledge to be evaluated. Since a semantic relation can be denoted by several textual patterns the quality of a triple can be extrapolated by: simply using IR measures to assign a value to its entities according to their co-occurrences in a collection of documents; or, based on the frequency of its entities connected by one or more patterns denoting its relation in the Web.

⋆ Financiado pela bolsa FCT SFRH/BD/44955/2008

This goal of this work is to analyse the benefits of applying measures based on the occurrence of words in documents to a set of relational triples. First, we start by using some well-known measures for weighting semantic triples extracted from Portuguese text. Then, considering the results from the first experiment, the similarity measures are adapted and, instead of looking for occurrences of the entities alone, they are used to search for these entities connected by textual patterns denoting semantic relations. Our assumption is that the scores given by the measures can be used to filter incorrect or less probable triples.

The work described here fits in a project [11] which has the goal of automatically constructing a lexical ontology for Portuguese.

This paper is organised as follows: in section 2 well-known IR similarity measures, are introduced; in section 3 we present the correlation values between the measures and manual evaluation; then, in section 4 we propose and assess an evaluation method; finally, before concluding (section 6), we refer some related work (section 5).

2 Similarity Measures

Among IR methods, we can find statistical approaches based on the occurrence of words in documents. Having in mind the distributional hypothesis [15], which assumes that similar words tend to occur in similar contexts, these methods are suitable, for instance, to find similar documents based on the words they contain or to compute the similarity of words based on their co-occurrence (see 2.1).

Another popular methods for computing the semantic similarity of words involves mathematical models based on the distribution of words in large corpora, such as the World Wide Web (see 2.2).

In this section, we present common similarity measures. Some of them are simple adaptations of popular co-occurrence measures. In their expressions:

- e_i and e_j correspond to entities, which can be words or expressions;
- $C = (d_1, d_2, ..., d_{|C|})$ is a collection of documents used to calculate the metrics;
- $|C|$ correspond to the number of documents contained by the collection C;
- $P(e_i)$ is the number of documents ($d_n \in C$) where e_i occurs;
- $P(e_1 \cap e_2)$ is the number of documents where e_i and e_j co-occur;
- and N is the total number of pages indexed in the corpus[1].

2.1 Corpus-based Similarity Measures

The measure of Cocitation (expression 1) was first presented in [23] as a similarity measure between scientific papers, after analysing their references. However, it has been applied to other contexts like the similarity between Web pages [7].

We have adapted this expression to measure the similarity between entities, where $P(e_i \cap e_j)$ is the number of documents containing both entities (e_i, e_j) and $P(e_i \cup e_j)$ is the number of documents containing at least one of the entities.

$$\mathrm{Cocitation}(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \quad (1) \qquad \mathrm{w}(e_i, d_j) = (1 + \log_2 f(e_i, d_j)) * \log_2 \left(\frac{|C|}{P(e_i)}\right) \quad (2)$$

[1] In case of Google, can be roughly estimated to 10^{10} [2].

TF-IDF (expression 2) is a popular measure in IR which weights (w) the relevance of a term (e_i) in a document (d_j), $w(e_i, d_j)$. Also, in the following expression, $f(e_i, d_j)$ is the frequency, or the number of times e_i occurs in d_j.

LSA [8] is a measure typically used to rank documents according their relevance to a query. Using this measure, higher ranked documents, which have higher cosine values, are those containing entities more similar to the query. In the calculation of LSA, the weight of each entity in a document ($w(e_i, d_k)$ and $w(e_j, d_k)$) can be obtained using TF-IDF, the number of occurrences of e_i in d_k, or other method to compute the relevance of a word in a document.

$$Lsa(e_i, e_j) = \frac{\sum_{k=1}^{|C|} w(e_i, d_k).w(e_j, d_k)}{\sqrt{\sum_{k=1}^{|C|} w^2(e_i, d_k)}\sqrt{\sum_{k=1}^{|C|} w^2(e_j, d_k)}} \quad (3)$$

Lin [18] proposes a measure which does not assume any kind of domain model as long as it has a probabilistic model and is not defined directly by a formula. Still, the measure is derived from a set of assumptions on similarity − the similarity between two objects is the ratio between the information common to both of the objects and the information needed to describe each one of them. Expression 4 is Lin's measure applied to the similarity of two terms, based on their distribution in a corpus. There, the information common to both terms is given by the documents where they co-occur and the information needed to describe them is the sum of the documents where each term occurs.

$$\mathrm{Lin}(e_i, e_j) = \frac{2 * \log P(e_i \cap e_j)}{\log P(e_i) + \log P(e_j)} \quad (4) \qquad \mathrm{Pmi}(e_i, e_j) = \log_2 \left(\frac{P(e_i \cap e_j)}{P(e_i) * P(e_j)} * |C| \right) \quad (5)$$

The algorithm called PMI-IR [24] uses Pointwise Mutual Information (PMI) to measure the similarity of pairs of words. More precisely, PMI-IR was used to identify (near) synonym words based on their co-occurrences in the Web, using expression 5, or variations of the latter, tuned for a specific search engine.

A completely different metric [17], based on the significance of the words in a corpus, was used to measure the similarity between two words. In expression 6, which measures the significance of entity e_i, the number of occurrences of e_i in corpus C is given by $O(e_i, C) = \sum_{j=1}^{N} f(e_j, C)$, where $O(e_i, C) \in N$. Expression 7 computes the similarity between entities e_i and e_j.

$$sim(e_i) = \frac{-\log\left(\frac{f(e_i, C)}{O(e_i, C)}\right)}{-\log\left(\frac{1}{O(e_i, C)}\right)} \quad (6) \qquad \sigma(e_1, e_2) = sim(e_1) * sim(e_2) \quad (7)$$

2.2 Web-based Similarity Measures

WebJ measure (expression 8) is an adaptation of the Jaccard coefficient.

$$\mathrm{WebJ}(e_1, e_2) = \frac{P(e_1 \cap e_2)}{P(e_1) + P(e_2) - P(e_1 \cap e_2)} \quad (8) \qquad \mathrm{WebO}(e_1, e_2) = \frac{P(e_1 \cap e_2)}{\min(P(e_1), P(e_2))} \quad (9)$$

The WebOverlap (WebO) (expression 9) and WebDice (WebD) (expression 10) measures are two variations of the WebJ, respectively for measuring the overlap and the mean overlap of two sets. More precisely, the Overlap minimises the effect of comparing two objects of different sizes, so the number of co-occurrences is divided by the lowest number of page counts, $\min(P(e_i), P(e_j))$.

$$\mathrm{WebD}(\mathrm{e}_1, e_2) = \frac{2 * P(e_1 \cap e_2)}{P(e_1) + P(e_2)} \quad (10) \qquad \mathrm{WebP}(\mathrm{e}_i, e_j) = \log_2\left(\frac{P(e_i \cap e_j)}{P(e_i) * P(e_j)} * N\right) \quad (11)$$

The WebP measure (expression 11) stands for PMI and quantifies the statistical dependence between two entities [24]. If entities e_1 and e_2 are statistically independent, the probability that they co-occur is given by $P(e_1) * P(e_2)$. On the other hand, if they tend to co-occur, $P(e_1 \cap e_2)$ will be higher than $P(e_1) * P(e_2)$, and the PMI will thus be greater.

The Normalised Web Distance (NWD, in expression 12) [5] measures the distance of two entities, based on their co-occurrences on the Web. Therefore, if the entities always co-occur, this means they are very similar and NWD is 0. On the other hand, although NWD most of the times ranges from 0 to 1, if the entities never co-occur, NWD is $+\infty$. For using NWD to obtain the similarity of two entities, we need to invert the distance and bound it to the [0-1] range [14], in a measure that we will call Normalised Web Similarity (NWS, expression 12).

$$NWS(e_1, e_2) = e^{-2*NWD(e_1, e_j)} = e^{-2*\frac{\max(\log P(e_1), \log P(e_2)) - \log P(e_1 \cap e_2)}{\log N - \min(\log P(e_1), \log P(e_2))}} \quad (12)$$

3 Using Similarity Measures to Weight Triples

This experiment analyses the benefits of applying metrics based on the occurrence of words and their neighbourhoods in documents to a set of relational triples automatically extracted from corpora. The extraction method uses the Onto.PT grammars, available though `http://ontopt.dei.uc.pt`.

3.1 Experiment set-up

Through this experimentation we have used the part-of-speech annotated version of the CETEMPúblico corpus [22], provided by Linguateca[2], containing text from the newspaper Público, published between 1991 and 1998.

We used only the first 28,000 documents of CETEMPúblico, which contain 30,100 unique content words (considering only nouns, verbs and adjectives) and results in approximately 1 million of *word-in-document* relations.

A relational database, which can be seen as an occurrence matrix, was used to store this information and also the TF-IDF of all words. This occurrence matrix provides: (i) the number of documents, d_k; (ii) the number of times the word w_i occurs, (iii) the documents where w_i occurs; (iv) the number of words in d_k, N_{d_k}; (v) the total number of words in the corpus, N; and (vi) the relevance R_{w_i} of the

[2] http://www.linguateca.pt

word w_i in the corpus. With this information we can calculate the co-occurrence between w_1 and w_2 and the number of times both occur $P(w_1 \cap w_2)$.

For experimentation purposes, the extraction method was also performed over the first 50,000 documents of CETEMPúblico and a total amount of 16,956 triples was obtained, more precisely 270 synonymy triples, 9,365 hypernymy, 1,373 part-of, 2,660 cause-of, and 3,288 purpose-of triples.

3.2 Application of the metrics

The distributional metrics referred in section 2.1, were implemented and normalised to fit the interval [0-100]. For instance, PMI-IR was normalised based on Bouma's [3] proposal. Also, calculation of the weights $w(e_i, d_k)$ in the LSA expression (3) was done by two different methods: the number of occurrences of entity e_i in the document d_k (LSA_o) and TF-IDF (LSA_t).

Each distributional metric was applied to the triple set, T, in the following manner: for each triple $t_i = (e_1, r, e_2), t_i \in T$, the distributional similarity between e_1 and e_2 was computed. For multi-word entities, the metrics were applied between each word of one entity and each word of the other, excluding stopwords, in order to calculate the average similarity value.

To evaluate the correlation of the results, we selected random samples for each type of relation. The samples' sizes took the type of relation into consideration and were used 1,156 triples, which were divided into ten random samples, each one evaluated by one of ten human judges.

Each human judge was asked to assign one of the following values to each triple, according to its quality: (0), if the triple was completely incorrect; (1), if the triple was not incorrect but something was missing in one or both of its arguments or the relation was very generic; or (2), if the triple was correct.

The samples contained several incorrect triples (e.g. 261 hypernymy triples were assigned with (0), 96 with (1) and 146 with (2)). Nevertheless, we were more interested in the correlation between the manual evaluation and the output values given by the metrics.

3.3 Manual evaluation versus Distributional measures

In order to observe the relationships between the manual evaluation and the output values given by the metrics, the correlation coefficients between them were computed and are shown in table 1. It is possible to observe that most metrics are strongly correlated with the quality of the triples, except for synonymy. This happens because all metrics, except σ, are based on co-occurrences and, in corpora text, synonymy entities, despite sharing very similar neighbourhoods, may not co-occur frequently in the same sentence [9] or even in the same document because they are alternative names for the same things. This might also be the reason for the low correlation coefficients with σ, which is based on the relevance of the terms.

Higher correlation coefficients are obtained for the hypernymy relation with the metrics of PMI, LSA and Cocitation, which suggests that hyponyms and their hypernyms tend to co-occur more frequently than causes or purposes. Also, there are more ways to denote the later relations in corpora text which led to less extracted and more incorrect triples. This is in conformity with an experience [12] where patterns denoting these relations were looked for in CETEMPúblico to validate semantic triples included in the lexical resource PAPEL[3]. On the other hand, part-of relations have good correlation coefficients with Lin's measure and LSA. Also, worth noticing that, the obtained values for LSA calculated with the occurrences of the entities (LSA_o) are very similar to the ones calculated with the TF-IDF (LSA_t). However, calculating the number of occurrences of a term in a document is much faster than computing the TF-IDF.

Table 1. Correlation coefficients between measures and manual evaluation.

Relation	**Cocitation**	**LSA_o**	LSA_t	**PMI**	Lin	σ
Hypernymy	0.58	0.58	0.58	0.46	0.41	0.08
Part-of	0.15	0.19	0.19	0.21	0.39	-0.22
Synonymy	-0.06	-0.02	-0.02	0.11	0.00	0.19
Causation	0.09	0.01	0.01	-0.01	-0.07	0.18
Purpose	0.02	0.09	0.07	0.18	0.00	0.10

4 Using the Web to Validate Triples

This experiment analyses how well the Web distributional measures suit the task of validating semantic relations. To this end, the measures presented in section 2.2, were adapted to quantify the similarity of the entities attached to textual patterns denoting semantic relations. Then, we performed an information retrieval task, where these measures are used to identify the correct triples.

4.1 Experiment set-up

We developed a system that implements all the measures, using Google as the *Web Search Engine*. Having in mind the validation of semantic relations, we followed previous hints [20] and adapted the measures to quantify the similarity between entities connected by patterns expressing semantic relations. This means that, for validating the triple $t = (e_1, r, e_2)$, we select a pattern indicative of relation r, π_{ri}, and do the following in the expressions of the measures:

- $P(e_1)$ is the number of search engine results for the query: "e_1 π_{ri}";
- $P(e_2)$ is the number of search engine results for the query: "π_{ri} e_2";
- $P(e_1 \cap e_2)$ is the number of search engine results for the query: "e_1 π_{ri} e_2".

Therefore, we created a set of indicative patterns for the relations we were validating, hyponymy and part-of. To increase the coverage of our patterns, we used not only those conveying the direct relation, but also the indirect.

Since semantic relations can be expressed by several different textual patterns, we could not select one best pattern. So, we decided to use two sets, Π_h

[3] In that experience, only 4% and 10% of causes and purposes were respectively confirmed, against 18% and 22% of confirmed hypernymy and meronymy instances.

and Π_p, consisting of the most frequent patterns for the studied relations, namely hyponymy and part-of. Furthermore, as the measures accept only one pattern at once, we computed the final scores by four distinct methods. Considering that Π_r has all the patterns for relation r, we sort a list, $S_m : |S_m| = |\Pi_r|$, containing the scores given by a similarity measure m, with each pattern $\pi_{ri} \in \Pi_r$, such that the best score is in S_{m1}. The final score is then given by:

- using simple co-occurrence, without including the patterns in the expressions (NP), used as a baseline;
- the score of the best pattern (B), S_{m1};
- the average of the scores given by the two best patterns (2B), $\frac{S_{m1}+S_{m2}}{2}$;
- the average of the scores given by all patterns (Av), $\frac{\sum_{i=1}^{|S_m|} S_{mi}}{|S_m|}$.

To this purpose, we have used WordNet 2.0[4] to collect two sets of hyponymy and part-of triples. In order to reduce noise due to ambiguities, we took advantage of the organisation of WordNet, which has the synsets ordered by the most frequent senses of the words and created the sets in the following way: we selected all the relation instances between synsets which are the first sense of their most frequent word, then for each of the latter, we defined relational triples held by the first word in the connected synsets. The final sets, contain 1100 hyponymy triples and 1100 part-of triples, set H and P respectively.

Sets H and P contained only correct triples, but regarding we needed incorrect triples, we created a third set, I, with 1010 random pairs of words, which we made sure to be not related by hyponymy or by part-of.

4.2 Identification of correct triples

In this experiment, we performed an information retrieval task, where the measures were used to filter incorrect triples automatically from our dataset. Besides the similarity measures (see section 2.2), we have used two simpler measures. One considers the number of page counts (nHits). The other marks the triple as correct if there is at least one page count for $P(e_1 \cap e_2)$, using no pattern (NP), one pattern (B), or two patterns (2B) (hasHits). According to the score returned by each measure, we tested several cut points (θ), and selected only the triples with a score higher than θ. Then, we computed the precision, recall and F_1 for measuring the quality and the quantity of triples selected.

For each measure, table 2 presents the best F_1 scores (in %) and the respective θ. These values are presented, first, using only the correct and incorrect triples ($C + I$), and then adding the triples with the wrong relations ($C + I + WR$), which, for this task, were considered to be incorrect. This way, we compare how the similarity measures behave ideally or in a more realistic scenario, where, sometimes, extracted triples only fail on identifying the type of the relation.

Even though some we had observe low correlations in our experiments, most of them achieve high F_1, and significantly improve the baseline. Yet, the best F_1 measure without the triples with the wrong relation is achieved by the WebP using the two best patterns, with $\theta = 16$ for hyponymy and $\theta = 33$ for part-of. On

[4] Available through `http://wordnet.princeton.edu`

Table 2. Best F_1 measures and respective θ.

	R		WebJ		WebO		WebD		WebP		NWS		hasHits
			F_1	θ	F_1	θ	F_1	θ	F_1	θ	F_1	θ	F_1
$C + I$	Hypo	NP	70.5	$1E^{-4}$	68.5	$1E^{-4}$	77.5	$2E^{-4}$	68.5	0	68.5	0	89.0
		B	94.9	$2E^{-4}$	68.5	0	95.4	$2E^{-4}$	96.2	26	96.1	0.15	96.0
		2B	94.1	$2E^{-4}$	68.5	0	95.1	$2E^{-4}$	96.3	16	96.0	0.05	92.6
		Av	86.3	$2E^{-4}$	68.5	0	90.9	$2E^{-4}$	96.2	3	91.5	0.05	-
	Part	NP	91.5	$2E^{-4}$	68.5	$1E^{-4}$	93.7	$2E^{-4}$	68.5	0	91.6	0.05	88.9
		B	93.9	$2E^{-4}$	80.7	$2E^{-4}$	94.0	$2E^{-4}$	94.3	32	94.7	0.25	92.8
		2B	93.8	$2E^{-4}$	75.6	0.05	93.8	$2E^{-4}$	94.7	33	94.9	0.2	94.1
		Av	86.7	$2E^{-4}$	68.5	0	90.3	$2E^{-4}$	94.5	4	87.1	0.05	-
$C + I + WR$	Hypo	NP	51.0	$1E^{-4}$	51.0	$1E^{-4}$	51.0	$1E^{-4}$	51.0	0	51.0	0	61.8
		B	75.3	$2E^{-4}$	54.7	$2E^{-4}$	75.3	$3E^{-4}$	69.9	28	75.2	0.25	69.5
		2B	74.8	$2E^{-4}$	51.1	0	74.9	$5E^{-4}$	69.9	16	74.1	0.25	68.2
		Av	72.7	$2E^{-4}$	51.1	0	75.1	$2E^{-4}$	71.9	4	74.8	0.05	-
	Part	NP	70.6	$2E^{-4}$	51.0	$1E^{-4}$	70.7	$4E^{-4}$	62.1	1	72.5	0.05	61.5
		B	65.5	$2E^{-4}$	62.1	$2E^{-4}$	65.2	$2E^{-4}$	86.9	42	62.3	0.2	65.7
		2B	65.4	$2E^{-4}$	59.4	0.05	65.6	$2E^{-4}$	85.5	41	67.5	0.2	68.8
		Av	59.7	$2E^{-4}$	51.0	0	62.3	$2E^{-4}$	68.6	4	60.7	0.05	-

the other hand, when the triples with the wrong relation are added, WebP is still the best for part-of triples. Using only the best pattern with $\theta = 42$, it achieves 86.9% F_1 scores. WebP is, however, worse for hyponymy, where WebJ, WebD and NWS outperform it in this order. In the latter scenario, all F_1 measures are lower, around 75%, which is still good, considering that, sometimes, if we change the type of a relation from hyponymy to part-of, we still get acceptable relations, as the following situation: *economy* hyponym-of *system* and *economy* part-of *system*.

5 Related Work

Classical IE systems rely on textual patters that frequently denote semantic relations (see [16]). However, some trade-off is often needed in the selection of patterns because, on the one hand, some of them occur rarely and, on the other hand, the most frequent are usually ambiguous. In order to increase the recall and to minimise the effort needed to encode the patterns, state of the art IE systems typically have a (weakly) supervised pattern learning component (e.g. [21]), which, nevertheless, is prone to extract more noise.

Therefore, some mathematical models have been proposed to filter incorrect triples [4] or to estimate the reliability of learned patterns [21]. This lead directly to higher precision and, eventually by using more ambiguous patterns, higher recall. Having in mind the distributional hypothesis [15], these filters are generally based on distributional similarity measures, which quantify the similarity of words according to their distribution in large corpora.

In the last decade, the Web became an attractive target for the extraction of huge quantities of knowledge (e.g. in [6] [1]). Furthermore, the Web became a very interesting "resource" for quantifying and validating knowledge extracted automatically, not only because of its size, variety of subjects and redundancy, but also because web search engines provide an efficient interface.

So, distributional similarity measures were adapted for the Web and were used, for instance, to rank semantic relations [6]. They have also been exploited as features and combined with lexical patterns indicating synonymy in a robust metric which was claimed [2] to outperform all web-based similarity metrics.

PMI-IR [24] is a popular measure for searching the Web for pairs of similar words. Variations of the PMI have been used to reduce the noise in information extracted from the Web [1]. However, in the latter works, the similarity measures were adapted and, instead of searching for the entities alone, they were used to measure the similarity between the entities and indicative patterns. The obtained scores can thus be exploited to assess the likelihood of the triples or the quality of the patterns.

More than assigning probabilities, the similarity measures can take advantage of the redundancy of the Web to validate knowledge, including not only semantic triples, but also question-answer pairs [19].

6 Concluding Remarks

We have performed several experiments to confirm if similarity measures are well suited to filter wrong or less probable and validate semantic relations. Firstly, we have shown that it is possible to use these measures to quantify triples according to the co-occurrence of their arguments in text. Then, we performed an information retrieval task consisting of the identification of correct triples, based on the scores of the measures, where all the measures had got F_1 scores. The results obtained are promising and we believe that the best performing measures can, be used as an alternative to manual evaluation of relational triples, extracted automatically from textual resources.

References

1. Blohm, S., Cimiano, P., Stemle, E.: Harvesting relations from the web: quantifiying the impact of filtering functions. In: Proc. 22nd National Conf. on Artificial Intelligence (AAAI). pp. 1316–1321. AAAI Press (2007)
2. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. In: Proc. 16th Intl. Conf. on the World Wide Web (WWW). pp. 757–766. ACM Press, New York, NY, USA (2007)
3. Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Proc. Biennial GSCL Conference 2009, Meaning: Processing Texts Automatically. pp. 31–40. Tbingen, Gunter Narr Verlag (2009)
4. Cederberg, S., Widdows, D.: Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: Proc. Conf. on Computational Natural Language Learning (CoNLL). pp. 111–118 (2003)
5. Cilibrasi, R., Vitanyi, P.M.B.: Normalized Web Distance and Word Similarity. Computing Research Repository, ArXiv e-prints (2009)
6. Cimiano, P., Wenderoth, J.: Automatic Acquisition of Ranked Qualia Structures from the Web. In: Proc. 45th Annual Meeting of the Association of Computational Linguistics (ACL). pp. 888–895. ACL Press, Prague, Czech Republic (2007)
7. Cristo, M., de Moura, E.S., Ziviani, N.: Link information as a similarity measure in web classification. In: Proc. 10th Symposium On String Processing and Information Retrieval. pp. 43–55. Springer (2003)

8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41, 391–407 (1990)
9. Dorow, B.: A Graph Model for Words and their Meanings. Ph.D. thesis, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart (2006)
10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998)
11. Gonçalo Oliveira, H., Gomes, P.: Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In: Proc. 5th European Starting AI Researcher Symposium (STAIRS'10). pp. 1135–1136. IOS Press (August 2010)
12. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Relations extracted from a Portuguese dictionary: results and first evaluation. In: Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA) (October 2009)
13. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação 2(1), 77–93 (May 2010), nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009
14. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In: Proc. 9th Intl. Conf. on Web Infor. Systems Engineering (WISE). pp. 136–150. Springer (2008)
15. Harris, Z.: Distributional structure. In: Papers in Structural and Transformational Linguistics, pp. 775–794. D. Reidel Publishing Company, Dordrecht, Holland (1970)
16. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. 14th Conf. on Computational Linguistics. pp. 539–545. ACL Press, Morristown, NJ, USA (1992)
17. Kozima, H., Furugori, T.: Similarity between words computed by spreading activation on an English dictionary. In: Proc. 6th Conf. on European Chapter of the Association for Computational Linguistics (EACL). pp. 232–239. ACL, Morristown, NJ, USA (1993)
18. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proc. 15th Intl. Conf. on Machine Learning. pp. 296–304. Morgan Kaufmann (1998)
19. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. In: Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 425–432 (2002)
20. Oliveira, P.C.: Probabilistic Reasoning in the Semantic Web using Markov Logic, pp. 67–73. University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering (July 2009)
21. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: Proc. 21st Intl. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL). pp. 113–120. ACL Press, Sydney, Australia (2006)
22. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for portuguese. In: Proc. 39th Annual Meeting on Association for Computational Linguistics (ACL'01). pp. 450–457. ACL, Morristown, NJ, USA (2001)
23. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science 24(4), 265–269 (1973)
24. Turney, P.D.: Mining the Web for Synonyms: PMI–IR versus LSA on TOEFL. In: Proc. 12th European Conf. on Machine Learning (ECML). vol. 2167, pp. 491–502. Springer (2001)