Discovering non-taxonomic relations from the Web

David Sánchez and Antonio Moreno

Universitat Rovira i Virgili (URV)
Computer Science and Mathematics Department
43007 Tarragona, Catalonia (Spain)
{david.sanchez,antonio.moreno}@urv.net

Abstract. The discovery of non-taxonomical relationships is one of the less studied knowledge acquisition tasks, even though it is a crucial point in ontology learning. We present an automatic and unsupervised methodology for extracting non-taxonomically related concepts and labelling relationships, using the whole Web as learning corpus. We also discuss how the obtained relationships may be automatically evaluated, using relatedness measures based on WordNet.

1 Introduction

The Web is an invaluable repository of knowledge. It has been considered that the number of resources available in the Web is so vast and the amount of people generating web pages is so enormous, that the Web information distribution approximates the actual real distribution as used in society [1]. Moreover, the redundancy of information in such a wide environment can represent a measure of relevance and trustiness of information for a certain domain. For those reasons, many authors [2][3][4] have been using the Web as the corpus for different knowledge acquisition tasks.

One of the most researched tasks is *ontology learning* from the Web. However, most of the approaches [4][5] are focused on the acquisition of taxonomical relationships and often neglect the importance of interlinkage between concepts. In fact, the discovery of non-taxonomic relations is understood as the least tackled aspect within ontology learning [6]. In general, two tasks have to be performed: first, detect which concepts are related and, second, assign a label for the relation (typically using verb phrases).

So, in this paper, we present an *automatic method for discovering non-taxonomic relationships from the Web*. This task involves the *i)* discovery of semantic patterns used for expressing non-taxonomic relationships in a specific domain (verb phrases) *ii)* retrieval of a corpus according to the acquired knowledge from where to extract candidates and *iii)* selection of the most appropriate concepts and relationships incorporating them to the ontology. This method has been designed as an extension of another one [7] that covers the taxonomical aspect of ontology learning.

The main features of our contribution are:

1. Unsupervised operation during the Web analysis and the learning process. This is important due to the amount of resources, avoiding the need of a human expert.
2. Automatic operation, allowing to perform easily executions to maintain the results updated. This characteristic fits very well with the dynamic nature of the Web.

3. Domain independent solution, because no domain assumptions are formulated and no predefined knowledge is needed. This is interesting when dealing with technological domains where specific and non widely-used concepts may appear. The only restriction here is that it can be only applied with English written resources.
4. Incremental learning with dynamic adaptation of the evaluated corpus as new knowledge is acquired. This results in an optimization of the computational cost of the analysis, retrieving only the most concrete and appropriate web resources.

The rest of the paper is organized as follows. Section 2 introduces the premises and techniques that configure the base of our proposal. Section 3 describes the proposed methodology to extract and label non-taxonomical relationships. Section 4 discusses the evaluation of results using WordNet based relatedness measures. Section 5 introduces related works and the final section presents the conclusions and lines of future work.

2 Knowledge acquisition framework

In this section, we comment some aspects of the knowledge acquisition techniques used to discover, extract and select non-taxonomical relationships from the Web:

- Lightweight analysis techniques are needed to achieve good scalability in such an enormous environment like the Web [8]. So, in order to perform an efficient analysis, the amount of processed information for each resource should be reduced to the minimum. Concretely, only those pieces of text that present knowledge in a simple, direct and unambiguous way (typically called *text nuggets* [8]) can be analysed.
- Statistical analysis applied over knowledge acquisition tasks is a good deal if enough information is available to obtain relevant measures. As introduced before, the Web environment is especially suitable for this task due to its size and heterogeneity. Moreover, web search engines can provide confident measures (*web-scale statistics*) in an immediate way if the appropriate search queries are formulated [9]. In our case, this is a crucial point because it allows us to obtain robust statistics about information distribution in a very scalable and efficient way.
- The use of linguistic patterns can be an effective technique to extract knowledge without expert's supervision and without predefined knowledge. For the taxonomical case, domain independent linguistic patterns [10] are a very common manner of discovering relationships. However, for the non-taxonomical case, aside from a reduced set of predefined relationships (e.g. meronymy, antonymy, synonymy, etc), there do not exist domain independent patterns, as those relationships are typically expressed by a verb that relates a pair of concepts [6]. If we want to use this pattern-based approach to extract non-taxonomical knowledge, a previous step for learning suitable patterns (verb phrases) for the analysed domain is required.
- As learning in a completely unsupervised and automatic way is difficult, an incremental approach in which several learning steps are defined and each one is enriched (bootstrapped) with the most relevant knowledge already acquired can be suitable. In our case, bootstrapping is used to constrain the search process contextualizing the queries formulated to the search engine in order to obtain appropriate corpus of documents and web-scale statistics for the specific domain. Bootstrapped information involves already acquired concepts and linguistic patterns.

3 Non-taxonomical learning

The first step in our non-taxonomical learning is the discovery of linguistic patterns that express non-taxonomical relationships. In this case, those relationships are typically expressed by a verb relating a pair of concepts. Due to the potential amount of verbs available in the English language, we should find which of them are truly relevant for the particular domain (in our case, a domain is expressed by an initial keyword, such us *hypertension*). In order to do this, we query a web search engine with the initial keyword to obtain a set of web resources covering the specific domain. For each one, a lightweight syntactical analysis is performed, considering the neighbourhood terms that surround the initial keyword in order to find verb phrases (conjugated verb and, optionally, prepositions), composing a list of candidates. Those candidates are classified in function of their position within the sentence: *predecessors* (e.g. *"is associated with* hypertension") or *successors* (e.g. "hypertension *is treated with*") of the initial keyword.

Each candidate is evaluated in order to decide if it is really closely related to the search domain. As we base our unsupervised analysis in statistical measures, we consider measures of co-occurrence between the verb phrase and the domain's keyword as a measure of relatedness between them. In order to obtain a robust measure, we use the mentioned web-scale statistics that represent the distribution of a queried concept in the whole Web. Concretely, for each verb phrase candidate that has been extracted as a *predecessor* of the initial keyword, we compute the following relatedness score (1) by asking the number of hits returned by a web search engine for the following queries:

$$Score(\frac{verbPhrase}{initKey}) = \frac{hits(\text{"}verbPhrase\ initKey\text{"})}{hits(\text{"}verbPhrase\text{"})} \quad (1)$$

Alternatively, if the candidate has been extracted as a *successor* of the initial keyword, we compute the same relatedness but specifying the inverse order in the corresponding query (*hits("initKey verbPhrase")*). Those formulas are derived from the score measures presented in [9] that are typically used to compute the degree of relationship between two words. Note also that double quotes are used to force the search of the exact string to ensure that the verb phrase is truly linked with the initial keyword.

The returned values are used to rank the list of domain dependent linguistic pattern candidates (verb phrases) and select those that are more closely related to the analysed domain (see examples in Table 1, for the *hypertension* example).

Once those domain dependent linguistic patterns have been obtained, the next step is to use them to discover concepts that are non-taxonomically (verb-labelled) related with the initial keyword. So, we query a web search engine with the pair "verb-phrase initial-keyword" or "initial-keyword verb-phrase" depending on whether the verb phrase precedes or succeeds the initial keyword. The result will be a set of web resources that contain the specified query. Our objective in this case is to evaluate their content to obtain concepts that immediately precede (e.g. "*salt intake* is associated with hypertension") or succeed (e.g. "hypertension is treated with *diuretics*") the specified query. Those new concepts become candidates for being non-taxonomically related with the initial keyword, labelling this relation with the verb phrase.

Next, we have to decide again which of the extracted concepts (e.g. "salt intake") are closely related to the searched domain (e.g. "hypertension"). In order to perform

Table 1. Firsts elements of the ranked list of verb phrases (173 total candidates) for the *Hypertension* domain, according to their position (PREdecessors or SUCcessors of the initial keyword).

Verb phrase	Position	Relatedness	Verb phrase	Position	Relatedness
is diagnosed in	SUC	0.12	is indicated for	PRE	0.11
are diagnosed as	PRE	0.10	is diagnosed as	PRE	0.08
is associated with	PRE	0.06	are associated with	PRE	0.06
is aggravated by	SUC	0.05	is cured by	SUC	0.03
is caused by	SUC	0.03	occurs during	SUC	0.03
is influenced by	SUC	0.03	suffer from	PRE	0.02
is treated with	SUC	0.02	accelerates	SUC	0.02

this selection process we use again web scale statistics about the co-occurrence of those two terms. In this case, the relatedness score is computed in the following manner (2):

$$Score(\frac{Concept}{initKey}) = \frac{hits(\text{"}initKey\text{"}\ AND\ \text{"}Concept\text{"})}{hits(\text{"}Concept\text{"})} \quad (2)$$

In this case, the AND operator ensures that those two terms co-occur within the text but not necessarily in the same sentence. If we use double quotes or add the verb phrase to the query, it will become too restrictive to obtain robust measures.

Those concepts whose relatedness is higher than a threshold (see some examples for *hypertension* in Table 2) are selected and incorporated into the ontology, with a relation that is labelled according to the verb phrase used to discover them (e.g. "salt intake" "is associated with" "hypertension"). Note that the direction of the relation corresponds to the role that each concept plays in the sentences (subject or object).

Finally, results are integrated with those from a methodology for learning taxonomies [7] using a standard ontology representation language (OWL), providing a tool that covers the main aspects of ontology learning.

Table 2. Examples of verb-labelled non-taxonomical relations for the *Hypertension* domain. Those in *italic* represent rejected candidates (relatedness below 0.1).

Subject (NP)	Verb (VP)	Object (NP)	Relat.
diuretic therapy	is indicated for	hypertension	0.61
salt intake	is associated with	hypertension	0.45
latter factors	*are associated with*	*hypertension*	*0.08*
hypertension	*is diagnosed in*	*individuals*	*0.006*
hypertension	is aggravated by	obesity	0.12
hypertension	*is aggravated by*	*the increase*	*0.01*
hypertension	is influenced by	sodium retention	0.65
hypertension	*is influenced by*	*some factors*	*0.02*
hypertension	is treated with	antihypertensives	0.55
hypertension	accelerates	renal disease	0.49
hypertension	*accelerates*	*the development*	*0.003*

4 Evaluation

Evaluating results of an automatic learning methodology is a difficult task. In general, either an expert opinion is needed to check the results manually or a repository is required to perform any automatic evaluation. Due to the nature of the learning environment (the huge and changing Web) and the learning method (domain independent, automatic and unsupervised) we focus on the automatic side.

The biggest and most widely used general purpose English electronic repository is WordNet [11]. It offers a lexicon, a thesaurus and semantic linkage between English terms. Using all this information it is possible to compute the similarity and relatedness between concepts. In this sense, the software *WordNet::Similarity* [12] offers an implementation of some standard measures that have been widely used for different knowledge related tasks [13]. Concretely, similarity measures use information found in an *is-a* hierarchy of concepts and quantify how much a concept *A* is like another *B*. WordNet is particularly well suited for similarity measures, since it organizes nouns and verbs into *is-a* hierarchies and, therefore, it can be very adequate for evaluating taxonomical relations. However, concepts can be related in many ways beyond being similar to each other. As such, WordNet provides relations beyond *is-a*, including *has-part*, *is-made-of* and *is-an-attribute-of*. In addition, each concept is defined by a short gloss that may include an example of use. All this information can be brought to bear in creating measures of relatedness. As a result, these measures tend to be more general and, in our case, more appropriate for evaluating non-taxonomically related terms.

Among the different relatedness measures implemented by *WordNet::Similarity*, we have chosen *vector-pairs* [13] because it does not depend on the interlinkage between words that, in many situations, has a poor coverage in the WordNet semantic network. This measure incorporates information from WordNet glosses as a unique representation for the underlying concept, creating a cooccurrence matrix from a corpus made up of the WordNet glosses. Each content word used in a WordNet gloss has an associated context vector. Each gloss is represented by a gloss vector that is the average of all the context vectors of the words found in the gloss. Relatedness between concepts is measured by finding the cosine between a pair of gloss vectors. For a pair of terms, the bigger the measure is, the more related the terms are (in a range between 0 and 0.5).

However, in general, all relatedness measures have serious limitations because they assume that all the semantic content of a particular term is modelled by semantic links and/or glosses in WordNet and, in consequence, in many situations, truly related terms obtain a low score due to WordNet's poor coverage. However, these measures are some of the very few fully automatic general purpose ways of evaluating results.

Applied to our approach, we check our Web based relatedness measure between two non-taxonomically related concepts by comparing it against *vector-pairs*, using *WordNet::Similarity* whenever both terms are in WordNet. The result can be represented in a plane in which each axis corresponds to one of those measures. Adding the limit that represents the selection threshold over both axis, it is also possible to evaluate the correctness of our candidate selection procedure visualizing correctly classified concepts (selected or discarded) and incorrectly classified ones (selected or discarded). An example of the type of results that we are able to obtain (whenever the particular domain is contained in WordNet) is represented in Figure 1 for the *Hypertension* domain.

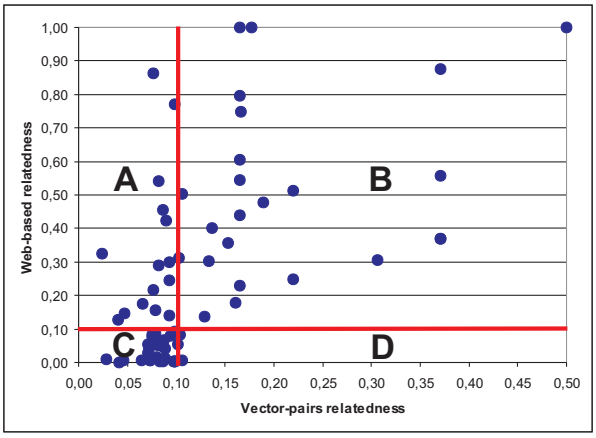

Fig. 1. Comparison of our Web-based relatedness measure against *vector-pairs* for the *Hypertension* domain. A: false positives; B: true positives; C: true negatives; D: false negatives.

Analysing this example (that shows a common tendency observed in several domains), we can extract the following conclusions:

- *True positives* (section B) cover a big area, with relatedness (using both measures) values above 0.1. *False negatives* (section D) are almost nonexistent apart from a pair of examples observed over the boundaries of the selection threshold. These facts indicate that when a pair of concepts are closely related according to the *vector-pairs* measure, our measure also indicates the same.
- *True negatives* (section C) tend to conform a compact set that have a value of relatedness (using both measures) below 0.1. However *false positives* (section A) are quite common, showing a discordance between both measures. Analysing them, we have observed that the poor performance is caused in many situations by the way in which *vector-pairs* (and in general all WordNet based relatedness measures) works. As has been introduced previously, those measures completely depend on WordNet's coverage for each concept (semantically expressed by interlinks or glosses); in consequence, when concepts are slightly considered in WordNet, those measures return a value that does not fully represent the reality. For example *vector-pairs* returns a very low value of 0.007 between *diuretics* and *hypertension* even though the first is a common treatment for the second; this is because, in Wordnet, this fact is not mentioned in the *diuretics*' gloss. In contrast, our measure depends on the Web's coverage for a particular term and, taking into consideration its size compared to WordNet, it can be seen why we are able to provide more consistent results over a wider set of concepts (returning a value of 0.6 for the mentioned example).

Finally, one may realize that the evaluation does not consider the verb used to label the relations. This is because relatedness measures are intended for nouns (concrete things with specific meaning). However, as final concepts are obtained through verb phrases, their quality (evaluated here) also depends on the quality of extracted verbs.

5 Related work

Faure and Nedellec [14] presented an interactive machine learning system, ASIUM, which hierarchically clusters nouns based on the verbs that they co-occur with and vice versa. The proposal by Byrd and Ravin [15] extracts named relations when they find particular syntactic patterns, such as an appositive phrase. They derive unnamed relations from concepts that co-occur by calculating the measure for mutual information between terms. Finkelstein and Morin [16] combine supervised and unsupervised extraction of relationships between terms, assigning them default labels. Maedche and Staab [17] use shallow text processing methods to identify linguistically related pairs of words. Thereby, they use the background knowledge from a taxonomy to propose relations at the appropriate level of abstraction but without considering the problem about labelling. Kavalec *et al.* [6] apply co-occurrence analyses to extract related terms. Then, they hypothesised that the 'predicate' of a non-taxonomic relation can be characterised by verbs frequently occurring in the neighbourhood of pairs of lexical entries. Other approaches based on clustering of documents using self-organizing maps (SOM) [18] or topological trees are able to express relationships between clusters unsupervisedly.

Studying those systems, the conclusion is that most of these approaches apply co-occurrences analysis in order to find out which concepts are related. In some cases, those unnamed relations are labelled using the verbs. Those aspects also conform the base of our approach, but the way in which we obtain resources to evaluate, the linguistic analysis and the computing of statistical measures are especially adapted to achieve good performance and efficiency in the Web environment.

6 Conclusions and future work

The presented methodology does not start from any predefined knowledge and, in consequence, it can be applied over domains that are not typically considered in semantic repositories. The automatic operation eases the updating of the results in highly changing domains without depending on a human expert. Those aspects conform an unsupervised and domain independent methodology for extracting non-taxonomical relationships from the Web.

As future lines of research, firstly, we can consider problems about semantic ambiguity presented in natural language resources. In this sense, complementary methods have been developed for dealing with polysemy [7] and synonymy [19] specially adapted to our working environment (web resources, search engines and lack of predefined knowledge) that we plan to integrate into our learning methodology.

Secondly, regarding the discovered verb labelled relationships, in order to obtain a computer understandable knowledge base that allows inference, verb labels should be interpreted (e.g. the verb phrase "*is included into*" expresses a "*part of*" type relationship). Standard classifications of verbs [20] could be used for this purpose, adding additional information about the semantic content of verb labelled relationships.

Finally, the proposed evaluation methodology should be improved in order to tackle the limitations described in Section 4. Maintaining the automatic operation (WordNet based), we plan to combine *vector-pairs* [13] with other additional relatedness measures [12] in order to improve the performance.

Acknowledgements. The work has been supported by *Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i del Fons Social Europeu* of Catalonia.

References

1. Cilibrasi, R., Vitanyi, P.: Automatic meaning discovery using Google (2004) Available at: http://xxx.lanl.gov/abs/cs.CL/0412098.
2. Brill, E., Lin, J., Banko, M., Dumais, S.: Data-intensive Question Answering. In: 10th Text Retrieval Conference. (2001)
3. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching very large ontologies using the WWW. In: Workshop on Ontology Construction (ECAI-00), Berlin, Germany (2000)
4. Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D., Yates, A.: Unsupervised named-entity extraction form the Web: An experimental study. Artificial Intelligence **165** (2005) 91–134
5. Cimiano, P., Staab, S.: Learning by Googling. SIGKDD **6** (2004) 24–33
6. Kavalec, M., Maedche, A., Skátek, V.: Discovery of Lexical Entries for Non-taxonomic Relations in Ontology Learning. In: SOFSEM'04. Volume 2932 of LNCS. (2004) 249–256
7. Sánchez, D., Moreno, A.: Automatic Generation of Taxonomies from the WWW. In: 5th International Conference on Practical Aspects of Knowledge Management. Volume 3336 of LNAI., Vienna, Austria (2004) 208–219
8. Pasca, M.: Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded. In: CICLing 2005. Volume 3406 of LNCS., Springer (2005) 280–292
9. Turney, P.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: 12th European Conference on Machine Learning, Germany (2001)
10. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: 14th International Conference on Computational Linguistics, France (1992) 539–545
11. Miller, G.: Wordnet: A lexical database. Communication of the ACM **38** (1995) 39–41
12. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity - Measuring the Relatedness of Concepts. In: 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Boston, USA (2004)
13. Patwardhan, S.: Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness (2003) Master of Science Thesis.
14. Faure, D., Nedellec, C.: Corpus-based conceptual clustering method for verb frames and ontology acquisition. In: LREC-98 Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, Granada, Spain (1998)
15. Byrd, R., Ravin, Y.: Identifying and extracting relations from text. In: 4th International Conference on Applications of Natural Language to Information Systems. (1999)
16. Finkelstein-Landau, M., Morin, E.: Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods. In: Workshop on Ontological Engineering on the Global Information Infrastructure. (1999)
17. Maedche, A., Staab, S.: Discovering Conceptual Relations from Text. In: 14th European Conference on Artificial Intelligence, Amsterdam, Netherlands (2000)
18. Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., Saarela, A.: Self organization of a massive document collection. IEEE Transactions on Neural Networks **11** (2000) 574–585
19. Sánchez, D., Moreno, A.: Automatic Discovery of Synonyms and Lexicalizations from the Web. In: Artificial Intelligence Research and Development. Volume 131. (2005) 205–212
20. Levin, B.: English Verb Classes and Alternations. Chicago University, Chicago, USA (1993)