Automatic discovery of synonyms and lexicalizations from the Web

David Sánchez [1] and Antonio Moreno

Department of Computer Science and Mathematics
University Rovira i Virgili (URV)

Abstract. The search of Web resources is a very important topic due to the huge amount of valuable information available in the WWW. Standard search engines can be a great help but they are often based only on the presence or absence of keywords. Thus problems regarding semantic ambiguity appear. In order to solve one of them, we propose a new method for discovering lexicalizations and synonyms of search queries based on a previously obtained taxonomy of terms for the specified domain.

Keywords. Web mining, web search, knowledge acquisition, information extraction

1. Introduction

The World Wide Web is the richest repository of information available. However, its size, heterogeneity and human oriented semantics suppose a serious obstacle in the search for the desired information. In this sense, Web search engines like Google have become a great help for the final user for accessing web resources. Although they do a great job in indexing a large amount of Web sites, their classification algorithms are very simple in the sense that they only check the presence or absence of a specific keyword, but are not able to analyse the semantic content of the web resources. That approach has some handicaps that are reflected on the final results returned to the user for a specific query.

More concretely, the user initiates a search by specifying a search keyword that is supposed to be significant for the desired domain. However, if several ways of expressing the same concept exist (e.g. *synonyms*, different *lexicalizations*, or even *morphological derivative forms*), those resources that are using those "alternative forms" for referring to the same domain will be omitted by the search engine (nowadays, only a few search engines incorporate stemming algorithms and/or perform meta-searches; lexicalizations, acronyms, and synonyms are not considered in any case). The user should perform different queries with different keywords in order to retrieve those sites (e.g. for the *Cancer* domain, you can use several "equivalent" forms for expressing the same concept like *cancer*, *carcinoma*, *neoplasm*...). Therefore, the detection of these terms is a fundamental

[1] Correspondence to: David Sánchez Ruenes, Department of Computer Science and Mathematics (DEIM). University Rovira i Virgili (URV). Avda. Països Catalans, 26. 43007. Tarragona. Spain. Tel.: +34 977 559681; Fax: +34 977 559710; E-mail: david.sanchez@urv.net.

task when using key-based web search approaches in order to explore exhaustively the corpus of web resources that really covers a knowledge domain.

So, in this paper we present a *new and unsupervised methodology for the discovery of lexicalizations and synonyms for a specific domain*. It uses a taxonomy of terms associated to the query, previously obtained through the method described in [10, 11] in order to perform the appropriate contextualization of the information. It also uses extensively a Web search engine to obtain the available web resources for the domain and perform several analysis. No other previous knowledge or user intervention are required to obtain reliable results. However, a semantic repository with synonym information is used to perform automatic evaluations of the results.

The rest of the paper is organised as follows. Section 2 summarizes the main approaches in this area. Section 3 introduces the characteristics of the taxonomy used as the starting point. Section 4 describes the novel approach for discovering lexicalizations and synonyms of terms. Section 5 describes the automatic evaluation procedure used to check the results. The final section contains the conclusions and proposes lines of future work.

2. Related work

There are several domain independent lexical databases that include synonym information, such as WordNet [4], BRICO [6], and EuroWordNet [14]. These systems ensure a certain level of quality, at the cost of a substantial amount of human labour. However, a major limitation of such lexicons is the relatively poor coverage of technical and scientific terms.

From a computer-based point of view, there are several methodologies that try to find lexicalizations and synonyms for a given keyword. Statistical approaches are based on co-occurrence of synonyms contexts [8]: synonyms are typically presented with similar sets of surrounding terms. A classical technique based on this idea is *Latent Semantic Analysis*. The underlying idea is that the aggregate of all the word contexts in which a given word appears, provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other [7]. However, these techniques tend to return closely related words but, sometimes, not truly "equivalent" ones [2].

Other techniques [13] identify different lexicalizations of concepts based on the assumption that they use a common set of ’core’ characters. These techniques can be useful for detecting lexical synonyms (alternative spellings such as *Pentium III*, *Pentium 3*, *Pent. 3*), but not for discovering semantic synonyms (e.g. *sensor* and *transducer*).

Recent approaches for synonymy detection [12] use the Web and, more concretely, web search engines to perform the selection of synonyms. Given a list of candidates for synonyms previously selected, they perform queries involving those candidates and their contexts into a web search engine to obtain statistics (number of hits) that measure word’s co-occurrence. In the same way as the firstly introduced techniques, those values of co-occurrence between candidate synonyms and specific contexts are considered as semantic similarity measures.

3. Taxonomy building methodology

As mentioned before, we use as a starting point a taxonomy of terms that are relevant for a specific domain. This hierarchy is built automatically through the methodology described in [10, 11] directly from the whole Web. The algorithm is based on analysing a large number of web sites in order to find important concepts for a domain by studying the *neighbourhood* of an initial *keyword*.

Concretely, in the English language, the immediate anterior word for a keyword is frequently *classifying* it (expressing a semantic specialization) [5]. So, this *previous word* is used for obtaining the taxonomical hierarchy of terms (e.g. *breast cancer* is a subclass of *cancer*). The process is repeated recursively in order to create deeper-level subclasses (e.g. *metastatic breast cancer* is a subclass of *breast cancer*), composing a hierarchy (see an example in Figure 1 for the *Cancer* domain).

The system relies on a search engine in order to search and access the available web resources from where to extract knowledge (concretely terms and taxonomical relations through a *previous word* analysis). It constructs dynamically the appropriate search queries for the search engine obtaining the most adequate corpus of web resources at each time. Moreover, the search engine is also used for checking the relevance of the extracted terms and evaluating the strength of the taxonomical relationships between them through a statistical analysis based on the number of estimated results available in the Web (web-scale statistics [12]).

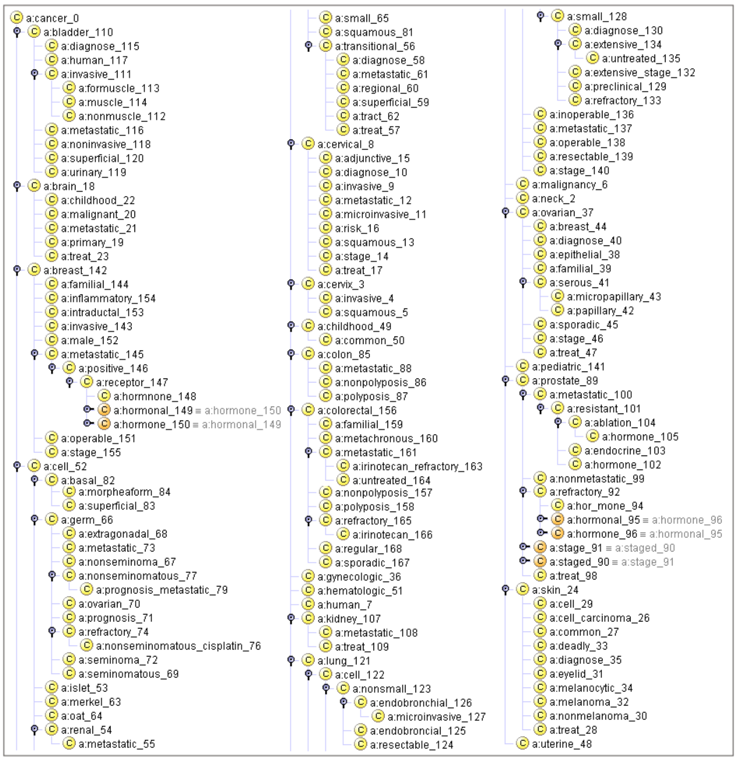

Figure 1. Example of obtained taxonomy for the *Cancer* domain.

4. Synonym and lexicalizations discovery

We have developed a novel methodology for discovering lexicalizations and synonyms using the taxonomy obtained by the construction algorithm for the given keyword and a web search engine. Our approach is based on considering the longest branches of subclasses (e.g. *hormone ablation resistant metastatic prostate cancer*) of the initial taxonomy and using them as the constraint (search query) for obtaining new documents that contain equivalent words for the main keyword. The assumption is that the longest multiword terms of the taxonomy contextualize enough the search to obtain, in most cases, lexicalizations or synonyms for the same semantic concept. Moreover, in order to check the relevance of the discovered candidates, statistics about co-occurrence with their contexts (multiword terms) in the domain are obtained from the number of hits returned by the search engine when formulating the appropriate query. These measures (web-scale statistics [12]) are very important due to the efficient and scalable way in which they can be obtained and their robustness, as they represent the presence of the query in the whole Web. Due to the Web's size and diversity, it can be assumed that these measures approximate the actual relative frequencies of those search terms as actually used in society [3].

In more detail, the discovery methodology works as follows:

- Select the N longest branches of the taxonomy, without considering the initial keyword (e.g. *hormone ablation resistant metastatic prostate*).
- For each one, make a query in the search engine of the whole multiword sentence and retrieve the first P pages. This can be made in two ways:

    1. Setting only the multiword term as the query (e.g. "*hormone ablation resistant metastatic prostate*"). That will return the webs containing this sentence without caring about the next word. Most of them will contain the original keyword but an amount will use lexicalizations or synonyms, ensuring that all pages will belong to the domain but slowing the search. This is the procedure followed for obtaining the results included in this section.
    2. Specifying the constraint not to contain the original keyword (e.g. "*hormone ablation resistant metastatic prostate*" *-cancer*). The set of pages (if there is any) will only contain alternative words. This will speedup the search dramatically but perhaps valid resources will be omitted (those that contain both the keyword and the alternative word(s)).

    In any case, a reduced set of web sites is enough to discover good candidates as the most suitable ones are typically found sooner than invalid ones, because the best synonyms are the ones that co-occur more frequently with their respective contexts, in this case, multiword's suffixes.
- Search among the text of the obtained web resources for the multiword term and evaluate the following word: the position that originally was occupied by the initial keyword (e.g. *cancer*). The word found position is considered to be a candidate for lexicalization or synonym (e.g. *carcinoma*).
- Repeat the process for each website and each multiword term and count the number of appearances of each candidate. A stemming morphological analysis is also performed for grouping different forms of the same word.

- Once the process is finished, a list of candidates is obtained. In order to select only the most reliable ones, a procedure to check their suitability based on statistical information retrieved from the web search engine is performed. For each candidate, a series of new queries to the web search engine using again multiword terms is performed in order to check if this candidate is commonly used as an alternative form for expressing the same concept in the domain, and not only with a few specific multiword terms from where it has been extracted. Concretely, for each multiword, a set of queries is constructed joining a suffix from that multiword and the new candidate. For example, for the *Cancer* domain, the *Carcinoma* candidate and the *hormone ablation resistant metastatic prostate* multiword, the domain constrained queries that could be performed are: "*prostate carcinoma*", "*metastatic prostate carcinoma*", "*resistant metastatic prostate carcinoma*", "*ablation resistant metastatic prostate carcinoma*" and "*hormone ablation resistant metastatic prostate carcinoma*". The longer the queries are, the more constrained and domain dependent they will be but, at the same time, the more difficult the obtaining of matching web sites will be. So, for example, queries of 2, 3 and 4 terms from each multiword (without counting the candidate) can be considered in this step. Each one is queried and the number of hits returned is considered. However, instead of evaluating the number itself (which will depend more on the generality of the multiword than on the candidate itself), we only consider the fact that the query has returned a minimum number (e.g. 10 hits). So, the number of queries that have returned some results is counted and weighted in function on the number of terms involved (1). If several derivative forms are available for the same candidate, the maximum relevance is considered.

$$relevance = \sum_{i=min\_terms}^{i=max\_terms} (i - min\_terms + 1) * \#queries\_with\_i\_terms \qquad (1)$$

- This final value represents the relevance of the candidate for becoming a final synonym or lexicalization for the domain, and allows selecting the most suitable ones (the more relevant, the closer to the domain the candidate is). As a refinement, it can be normalised in function on the number of total possible queries (2), obtaining a final percentage that eases the selection process (establishing a minimum threshold). In addition, with this measure, it is easy to detect and directly discard misspelled candidates as they typically return zero values.

$$relat\_relev = \frac{relevance}{\sum_{i=min}^{i=max} (i - min\_term + 1) * \#tot\_queries\_i\_terms} * 100 \qquad (2)$$

The described methodology has been tested with several domains obtaining promising results. For illustrative purposes, in tables 1, 2 and 3, results for the *Cancer*, *Disease* and *Sensor* domains respectively are presented.

Table 1. Firsts and lasts elements of the sorted list of lexicalizations and synonyms candidates for the *Cancer* domain (64 total candidates). From the obtained taxonomy, 31 multiwords of 3 terms and 16 multiwords of 4 terms have been considered evaluating 100 web sites including the original keyword. Elements in **bold** represent correctly selected results (see evaluation procedure in section 4)

Concept (root)	Derivatives	Relevance	Relative relev.	Correct?
cancer	**cancer, cancers**	**61**	**96,82%**	**true**
carcinoma	**carcinoma, carcinomas**	**30**	**47,62%**	**true**
tumor	**tumor, tumors**	**25**	**39,68%**	**true**
tumour	**tumours, tumour**	**24**	**38,09%**	**true**
neoplasm	**neoplasms**	**7**	**11,11%**	**true**
testi	testis	6	9,52%	false
bladder	bladder	5	7,93%	false
malign	malignancies, malignant	3	4,76%	false
epithelioma	**epitheliomas**	**2**	**3,17%**	**true**
carcino	carcino	2	3,17%	-
…	…	…	…	…
tumorsovarian	tumorovarian	0	0%	-

Table 2. Firsts and lasts elements of the sorted list of lexicalizations and synonyms candidates for the *Disease* domain (127 total candidates). From the obtained taxonomy, 84 multiwords of 3 terms and 24 multiwords of 4 terms have been considered evaluating 100 web sites including the original keyword. Elements in **bold** represent correctly selected results (see evaluation procedure in section 4)

Concept (root)	Derivatives	Relevance	Relative relev.	Correct?
diseas	**disease, diseases**	**122**	**92,24%**	**true**
disord	disorder, disorders	17	12,87%	false
syndrom	syndrome, syndromes	13	9,84%	false
lesion	**lesions**	**7**	**5,3%**	**true**
condit	**condition, conditions**	**7**	**5,3%**	**true**
stenosi	**stenosis**	**7**	**5,3%**	**true**
atherosclerosis	atherosclerosis	6	4,54%	false
infect	**infections, infection,infectivity, infects**	**6**	**4,54%**	**true**
stenos	**stenoses**	**6**	**4,54%**	**true**
obstruct	obstruction, obstructions	5	3,78%	false
…	…	…	…	…
diseaseinform	diseaseinformation	0	0%	-

Table 3. Firsts and lasts elements of the sorted list of lexicalizations and synonyms candidates for the *Sensor* domain (27 total candidates). From the obtained taxonomy, 17 multiwords of 3 terms and 1 multiword of 4 terms have been considered evaluating 100 web sites including the original keyword. Elements in **bold** represent correctly selected results (see evaluation procedure in section 4)

Concept (root)	Derivatives	Relevance	Relative relev.	Correct?
sensor	**sensor, sensors, sensores**	**17**	**89,47%**	**true**
transduc	**tranducer, transducers**	**4**	**21,05%**	**true**
measure	measurement	2	10,5%	false
circuit	**circuit**	**2**	**10,5%**	**true**
signal	signal	2	10,5%	false
…	…	…	…	…
code	codes	0	0%	false

5. Evaluation

In order to perform an evaluation of the results in an automatic and domain-independent way, we are using WordNet [4]. WordNet offers a lexicon, thesaurus and semantic linkage between the major part of English terms.

In our case, we perform an automatic evaluation of the obtained results against the synsets (synonym sets) presented in WordNet for a specific keyword in comparison to our list of sorted candidates. In WordNet, each synset groups a set of concepts that are considered to be truly equivalent, and assigns them a gloss. However, due to the proliferation of a high number of unclear word sense distinctions [1] and the subtle semantic organization of terms, in many situations synsets are quite incomplete (e.g. *Disease* has not got any synonym). However, as our final purpose for synonyms discovery is to widen the search process using other typically equivalent forms for expressing the same concept, other semantic related terms can be also considered. Concretely, first levels of *hyponym* or *hypernym* terms for a specific concept are typically used as equivalent terms (e.g. *cancer* is a hypernym of *carcinoma*).

Taking these facts into consideration, the automatic evaluation procedure can be performed in the following form: for each discovered candidate that is included in WordNet, the number of semantic links between it and the original concept following *hyponym* and/or *hypernym* pointers is computed; those that present a semantic distance close enough (4 pointers maximum in our case) are considered to be correctly selected as final synonyms (see the last column in tables 1, 2 and 3). Although this process is automatic, the procedure can only be considered as a first approximation for evaluation because the semantic linkage of WordNet is far from complete or exhaustive enough especially in scientific and technological domains [12]; as a consequence, in some cases, suitable candidates are not considered (e.g. *disease* and *syndrome*). In the future, other more accurate ways for computing the semantic similarity between terms considering other semantic relationships included in WordNet (e.g. *meronyms, similar to*, *attribute*, etc) can be also considered [9].

6. Conclusions and future work

Taking into consideration the amount of resources available on the Web and its growing factor, we believe that methodologies that ease the search of information should be developed. Standard search engines are widely used for this task but they present serious limitations because their pattern search algorithms lack any kind of semantic content. Concretely, in order to extend the search and retrieve the largest amount of resources that are *semantically* relevant for the specified query, an algorithm for discovering alternative keywords (*lexicalizations* and *synonyms*) for the domain is proposed. This is useful for domains with a little amount of available web resources.

It is important to note that the proposed methodology performs in an automatic, domain independent and unsupervised way. Moreover, the use of Web scale statistics to check the relevance of candidates results in a highly efficient, scalable and robust solution.

As future lines of research, we plan to extend the lexicalizations and synonyms discovery to any sublevel of the taxonomy (not only to the initial keyword) covering even groups of words (e.g. *blood cancer* could be equivalent to *leukaemia* and *B.C.* could be an abbreviation of *breast cancer*). Moreover, the evaluation procedure will be improved considering other metrics for computing more accurately the semantic distance between concepts. From the taxonomical point of view, those discovered terms could be very useful for improving or extending the initial taxonomy (by searching in a new corpus of relevant documents that cover the same topic), or even to check the consistency of the hierarchy of terms, detecting classes and relationships that are maintained for different synonyms.

Acknowledgements

This work has been supported by the "*Departament d'Universitats, Recerca i Societat de la Informació*" of Catalonia.

References

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez: Enriching very large ontologies using the WWW. *In Proceedings of the Workshop on Ontology Construction of the European Conference of AI*, 2000.

[2] V. Bhat, T. Oates, V. Shanbhag and C. Nicholas: Finding aliases on the web using latent semantic analysis. *Data Knowledge & Engineering* **49** (2004), 129-143.

[3] R. Cilibrasi and P.M.B. Vitanyi: Automatic meaning discovery using Google. http://xxx.lanl.gov/abs/cs.CL/0412098, 2004.

[4] C. Fellbaum: WordNet: An Electronic Lexical Database. Cambridge, Massachusetts, 1998.

[5] G. Grefenstette: The World Wide Web as a resource for example-based Machine Translation Tasks. *In Proceedings of Aslib Conference on Translating and the Computer*, London, 1999.

[6] K. Haase: Interlingual BRICO. *IBM Systems Journal* **39** (2000), 589-596.

[7] T.K. Landauer and S.T. Dumais: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review* **104** (1997), 211-240.

[8] C.D. Manning and H. Schütze: Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts, 1999.

[9] T. Pedersen, S. Patwardhan and J. Michelizzi: WordNet::Similarity - Measuring the Relatedness of Concepts. American Association for Artificial Intelligence, 2004.

[10] D. Sánchez and A. Moreno: Creating ontologies from Web documents. *In Proceedings of the Setè Congrés Català d'Intel.ligència Artificial*, IOS Press **113** (2004), 11-18.

[11] D. Sánchez and A. Moreno: Automatic generation of taxonomies from the WWW. *In Proceedings of the 5th International Conference on Practical Aspects of Knowledge Management*, LNAI **3336** (2004), 208-219.

[12] P.D. Turney: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *In Proceedings of the Twelfth European Conference on Machine Learning*, 2001.

[13] A.G. Valarakos, G. Paliouras, V. Karkaletsis and G. Vouros: Enhancing Ontological Knowledge Through Ontology Population and Enrichment. *In Proceedings of EKAW*, LNAI **3257** (2004), 144-156.

[14] P. Vossen: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht, Netherlands: Kluwer. See: http://www.hum.uva.nl/ ewn/, 1998.