Development of new techniques to improve Web search[*]

David Sánchez, Antonio Moreno
Department of Computer Science and Mathematics
University Rovira i Virgili (URV)
Avda. Paїsos Catalans, 26. 43007 Tarragona (Spain)
{david.sanchez, antonio.moreno}@urv.net

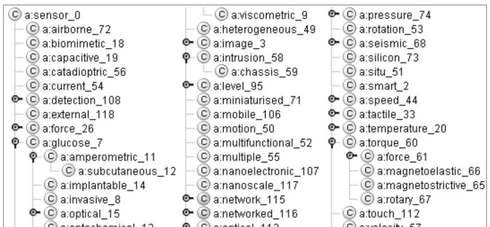
Abstract

Web search engines are a great help for accessing web sites but they present several problems regarding semantic ambiguity. In order to solve them, we propose new methods for polysemy disambiguation of web resources and discovery of lexicalizations and synonyms of search queries.

1 Introduction

Search engines like Google have become an indispensable tool. However, their ranking algorithms are based only on a keyword's presence and have some serious handicaps that are reflected on the final results. On the one hand, if several ways of expressing the same search query exist (e.g. *synonyms*, *lexicalizations*, or even *morphological forms*), a considerable amount of relevant resources will be omitted. On the other hand, if the user selects a polysemic word as a query, the set of returned results will contain sites that are using that concept in different contexts (e.g. *Virus*).

In order to face these two problems, in this paper we present *new, automatic and autonomous methodologies for semantic disambiguation and classification of web resources and for discovery of lexicalizations and synonyms for a specific domain.* These techniques use a previously obtained taxonomy of terms and web resources through the method described in [Sanchez and Moreno, 2004] to perform the appropriate contextualization (see an example for the *Sensor* domain in figure 1).



```
a:sensor_0
  a:airborne_72
  a:biomimetic_18
  a:capacitive_19
  a:catadioptric_56
  a:current_54
  a:detection_108
  a:external_118
  a:force_26
  a:glucose_7
    a:amperometric_11
      a:subcutaneous_12
    a:implantable_14
    a:invasive_8
    a:optical_15
    a:optochemical_13
    a:subcutaneous_10
      a:viscometric_9
  a:heterogeneous_49
  a:image_3
  a:intrusion_58
    a:chassis_59
  a:level_95
  a:miniaturised_71
  a:mobile_106
  a:motion_50
  a:multifunctional_52
  a:multiple_55
  a:nanoelectronic_107
  a:nanoscale_117
  a:network_115
  a:networked_116
  a:optical_113
    a:multibeam_114
  a:pressure_74
  a:rotation_53
  a:seismic_68
  a:silicon_73
  a:situ_51
  a:smart_2
  a:speed_44
  a:tactile_33
  a:temperature_20
  a:torque_60
    a:force_61
    a:magnetoelastic_66
    a:magnetostrictive_65
    a:rotary_67
  a:touch_112
  a:velocity_57
  a:wireless_119
```

Figure 1. Example of an initial taxonomy of *Sensors*.

2 Discovery of lexicalizations and synonyms

There are several methodologies that try to find lexicalizations for a keyword like *Latent Semantic Analysis* [Deerwester *et al*, 1990]. It considers that words that co-occur in the same documents are semantically related. However, it tends to return domain related words but, sometimes, not truly "equivalent" ones [Bhat *et al*, 2004]. Other techniques [Valarakos *et al*, 2004] identify lexicalizations with the assumption that they use a common set of 'core' characters. They can detect alternative spellings (e.g. *Pentium III*, *Pent. 3*), but not synonyms (e.g. *sensor* and *transducer*).

We have developed a novel methodology for discovering lexicalizations and synonyms using an initial taxonomy and a web search engine. Our approach is based on considering the longest branches (e.g. *subcutaneous amperometric glucose sensor*) of the taxonomy as a contextualization constraint and using them as the search query ommiting the initial keyword (e.g. "*subcutaneous amperometric glucose*") for obtaining new webs. In some cases, those documents will contain equivalent words for the main one just behind the searched query that can be considered candidates for lexicalizations or synonyms (e.g. *subcutaneous amperometric glucose transducer*). For each candidate (see Table 1 with candidates for the *Sensor* domain), a relevance measure (1) is computed, obtaining a ranked candidate list. Those that overpass a minimum threshold computed in function on the number of multiwords considered will be selected.

$$relev = \#Dif\_Web\_Appear * (10\hat{}(\#Multi\_words\_terms - 1)) \qquad (1)$$

Table 1. Lexicalizations and synonyms candidates for the *Sensor* domain (24 multiwords). Elements in **bold** are selected results.

Concept and derivates	Hits	Webs	Multi	Relev.
sensor, sensors	**437**	**294**	**24**	**2E25**
transducer, transducers	**28**	**19**	**3**	**1900**
system	4	4	2	40
probe	2	2	2	20
transmitter	11	11	1	11
measurement	10	7	1	7

[*] The support of "*Departament d'Universitats, Recerca i Societat de la Informació*" of Catalonia is acknowledged.

3 Polysemy detection and semantic clustering

In this case our goal is to be able to group different concepts (and their associated web sites) of the taxonomy associated to a search query in sets according to the concrete sense in which they are used. So, the user will obtain disambiguated sets of relevant web sites for the desired domain.

For that purpose, many techniques [Ide and Véronis, 1998] use an external dictionary to select the most appropriate meaning in each case. However, performing this classification without previous semantic knowledge is not a trivial process [Sanderson and Croft, 1999]. In this sense, we can take profit from the context where each concept has been extracted, concretely, the web documents that contain it. We can assume that each website included in the taxonomy is using the initial keyword (e.g. *virus*) in a concrete sense, so concepts that are selected from the analysis of that single document (e.g. *linux*, *email*) belong to the domain associated to the same keyword's meaning. Applying this idea over a significant amount of web sites we can find, as shown in figure 2 for the *Virus* example, quite consistent semantic relations between concepts.

The algorithm performs a clusterization process with the list of terms of a specific level of the taxonomy. For each pair, a similitude measure (2) is computed based on the amount of coincidences between the concept's URL sets. The resulting similarity matrix allows us to detect the most similar concepts and join them. A new matrix is computed for the resulting set of classes. This process is repeated until no more concepts can be joined (there are no coincidences between URLs), building a dendogram (see figure 2).

$$sim(A,B) = Max\left(\frac{\#Coin(URL(A),URL(B))}{\#URL(A)}, \frac{\#Coin(URL(A),URL(B))}{\#URL(B)}\right) \qquad (2)$$

The result is a partition (with 2 elements for the *virus* domain as shown in figure 2) of classes that groups the concepts and web resources that belong to a specific meaning. The number of final classes is automatically discovered by the clustering algorithm thanks to the constrained joining criteria. Note that this methodology can be applied to a set of terms at any level of the taxonomy.

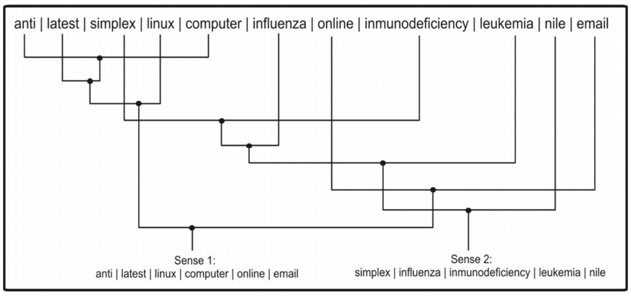

anti | latest | simplex | linux | computer | influenza | online | inmunodeficiency | leukemia | nile | email
Sense 1:
anti | latest | linux | computer | online | email
Sense 2:
simplex | influenza | inmunodeficiency | leukemia | nile

Figure 2. Dendogram representing semantic associations between the classes found for the keyword "*virus*". Two final clusters are obtained: Sense 1 groups the classes associated to "*computer program*" and Sense 2 the ones related to "*biological agent*".

4 Conclusions

Taking into consideration the amount of resources available easily on the Web, 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 due to their pattern-search algorithms because they lack any kind of semantic content.

On the one hand, *polysemy* becomes a serious problem when the retrieved resources obtained from the search engine are only based on the keyword's presence; in this case, we propose an automatic approach for *polysemy disambiguation of concepts and web resources*. On the other hand, in order to extend the search widely and retrieve the largest amount of resources that are *semantically* relevant for the query specified, an *algorithm for discovering alternative keywords* (*lexicalizations* and *synonyms*) for the same domain is also proposed. This is very useful for domains with a little amount of available web resources.

Moreover, it is important to note that the proposed methodologies perform in an *automatic* and *autonomous* way, allowing to maintain the results updated easily by performing executions as frequently as desired. This is a very useful feature when dealing with a highly changing environment like the Web and distinguishes our approach from other similar ones in which user's interaction or a thesaurus like WordNet are required [Lamparter *et al*, 2004].

References

[Bhat *et al*, 2004] V. Bhat, T. Oates, V. Shanbhag and C. Nicholas: Finding aliases on the web using latent semantic analysis. *Data Knowledge & Engineering* 49, 2004.

[Deerwester *et al*, 1990] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman: Indexing by latent semantic analysis, *Journal of the American Society of Information Science* 41 (6) pp. 391-407. 1990.

[Ide and Véronis, 1998] N. Ide and J. Véronis: Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics* 24. 1998.

[Lamparter *et al*, 2004] S. Lamparter, M. Ehrig and C. Tempich: Knowledge Extraction from Classification Schemas. ODBASE 2004, LNCS 3290, 2004.

[Sanchez and Moreno, 2004] D. Sanchez and A. Moreno: Automatic Generation of Taxonomies from the WWW. In: *Proceedings of the 5th International Conference on Practical Aspects of Knowledge Management*. LNAI, vol. 3336, Vienna, 2004.

[Sanderson and Croft, 1999] M. Sanderson, B. Croft: Deriving concept hierarchies from text. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1999, Berkeley, USA.

[Valarakos *et al*, 2004] A.G. Valarakos, G. Paliouras, V. Karkaletsis and G. Vouros: Enhancing Ontological Knowledge Through Ontology Population and Enrichment, EKAW 2004, LNAI 3257, 2004.