Universitat Oberta de Catalunya

DIGITHUM

The Humanities in the Digital Era

http://digithum.uoc.edu

Dossier "Academic research into Wikipedia"

Using Wikipedia to develop language resources: WordNet 3.0 in Catalan and Spanish

Antoni Oliver
Lecturer of the Department of Arts and Humanities and director of the postgraduate course on Translation and Technologies (UOC)
aoliverg@uoc.edu

Salvador Climent Roca
Lecturer of the Department of Arts and Humanities (UOC)
scliment@uoc.edu

Submission date: March, 2012
Accepted date: April, 2012
Published in: May, 2012

Abstract

We describe the state of the art in the use of Wikipedia for natural language processing tasks and also describe three applications of our own that enrich a powerful language resource: WordNet version 3.0 in Catalan and Spanish. Researchers have for many years sought applications that would take account of world knowledge in a more or less structured way, as this kind of knowledge has proven to be crucial to satisfactorily solving certain language processing tasks. Wikipedia may be the answer to the provision of this kind of information, as it is constantly updated and access is free.

Keywords

Wikipedia, WordNet, Natural Language Processing, linguistic resources

Ús de la Viquipèdia per al desenvolupament de recursos lingüístics: el WordNet 3.0 català i castellà

Resum

En aquest article presentem l'estat de la qüestió en l'ús de la Viquipèdia per a tasques relacionades amb el processament del llenguatge natural i tres aplicacions que hem creat per a l'enriquiment d'un recurs lingüístic de gran abast: el WordNet versió 3.0 per al català i castellà. Els investigadors en aquesta àrea fa anys que cerquen vies perquè les aplicacions integrin informació sobre coneixement del món, d'una manera més o menys estructurada, ja que aquest tipus de coneixement ha demostrat ser molt important per a resoldre de manera satisfactòria moltes tasques de processament del llenguatge. La Viquipèdia pot respondre perfectament a aquesta demanda d'informació amb l'avantatge del seu accés lliure i la seva actualització constant.

Paraules clau

Viquipèdia, WordNet, processament del llenguatge natural, recursos lingüístics

Introduction

A key goal of artificial intelligence (AI) is to equip machines with the ability to understand human language. Whereas language processing at the morphological and syntactic levels has reached a satisfactory stage in development, a major challenge is for computers to understand and represent meaning in order to be able to intelligently complete tasks.

Linguistic theory, following on from Langacker (1987), postulates that semantic processing cannot be considered to be linguistic but also encyclopedic, that is, word knowledge is inseparable from world knowledge. From this perspective, words are points of access to vast repositories of related knowledge. Many natural language processing (NLP) and AI researchers adopt the same starting point and seek methods and information sources that enable use of the knowledge collected and organized by humans.

In the last decade, Wikipedia (WP) has developed into a major semantic knowledge resource for NLP. WP is a vast, reliable and structured knowledge corpus that contains far more knowledge than any other manually created ontology or knowledge base, such as the extensive CYC (Lenat *et al.*, 1990) or the widely used WordNet (Fellbaum, 1998). WP also has all the NLP utilities of the typical corpus, primarily in regard to how it makes use of statistical methods. In addition, the simple and well-structured XML format and an ongoing dumping policy (backups of packages of all of WP up to a specific cutoff point) both greatly simplify processing.

WP's structure facilitates automatic and semiautomatic exploitation of its knowledge. WP is composed of articles organized as an article per concept and a concept per article. Titles are representative and introductions, relevant compact descriptions and infoboxes all reflect related knowledge structures. There are redirect links between equivalent terms, hyperlinks between terms and other articles, disambiguation pages for homonyms and synonyms and an underlying category structure by means of which each article is assigned to one or more categories. Wikipedia is also largely multilingual, with connected up structures for the different languages. WP articles may also have links to external URLs.

Hence, for the first time, NLP has available to it vast quantities of quality knowledge for exploitation. Below we review the current state of research in this area and then describe multilingual strategies of our own for extracting monolingual semantic knowledge.

1. The state of the art

Thomas and Amit (2007) have demonstrated how the quality of the information in WP articles can be inferred from the editing history: there is a clear correlation between stability in terms of production changes and WP quality, which is considered to be comparable to the Encyclopaedia Britannica (cf. Giles, 2005), just ten times larger (information referring to the year 2009).

Below we describe the main uses being made of WP for NLP research purposes, drawing especially on work reported in Medelyan *et al.* (2009), Gabrilovich *et al.* (2009) and Vivaldi *et al.* (2010).

1.1. Lexical disambiguation, semantic similarity and automatic document classification

Reasoning about the similarity between two words, phrases or documents is a routine task for humans but a challenge beyond the scope of machines. The most common way to establish the similarity between two documents is to statistically analyse the degree of coincidence between words, weighted according to the occurrence of these same words in text corpora; this approach is called latent semantic analysis (LSA) (Deerwester *et al.*, 1990). Gabrilovich *et al.* (2007) also developed a technique for measuring the degree of association between the words in a particular document and WP entries, with the comparison producing an estimate of similarity that has proved more accurate than the LSA.

Calculating similarity is potentially useful for many tasks, such as detecting plagiarism and automatically classifying documents; the latter task, in turn, can be used to route news and emails and to identify spam.

A limiting factor in reasoning regarding text string similarity is disambiguating lexical meaning, that is, deciding whether two words that are identical in form have the same meaning. This is a key task in all applications for recovering, extracting and processing information. It is also the most difficult task, because language is inherently ambiguous: humans constantly disambiguate between possible meanings according to the context in which words occur. For NLP purposes, WordNet (Fellbaum, 1998) is the knowledge base most used as an inventory of words with meanings organized as related concepts. However, this knowledge base has drawbacks, namely, the excessive proliferation of meanings, bias towards the common lexicon and low content in proper names and specialized terms.

Authors such as Medelyan *et al.* (2008) use techniques such as those described above to calculate similarity between documents, comparing concepts (represented by WP entries) that are candidate meanings for the analysed word with the words that constitute the context for the sentence in which the word appears.

WP is currently the largest known repository of proper nouns and specialized vocabulary. Searches link words in texts with WP articles, thereby disambiguating and categorizing words; in a certain context, for example, "Chicago" is the band and not the city and not the Broadway show.

Another linguistic task necessary for all intelligent applications is knowing when different words in a text refer to the same concept, eg., when "the King" is synonymous with "Juan Carlos

I". To solve this problem and so determine meaning, as well as applying the similarity measures described above, authors such as Ponzetto *et al.* (2006) exploit the appearance of words in the first paragraph of the WP article along with the article's links and the lists of common WP categories.

Automatic document classification organizes collections of documents into predefined categories on the basis of the words they contain. This simple approach functions poorly, however, when different documents refer to the same topics using different words. Gabrilovich *et al.* (2006) and other researchers optimize this task by growing the word lists for documents with related concepts extracted from WP. Banerjee (2007) notes that this strategy makes the method more adaptable as it provides information with more stable foundations that are independent of ongoing changes in terminology.

1.2. Retrieving information and obtaining answers

Retrieving information and obtaining answers from collections of documents is a task that is reiteratively performed nowadays using popular search engines to pose questions and using the Internet as a collection of documents. WP is used to improve this approach in different ways.

First, search questions can be automatically expanded in WP using synonyms, alternative forms and related concepts so as to retrieve relevant documents. These operations had already been implemented in WordNet but are optimized with WP (eg., Gregorowicz *et al.*, 2006); this is due to the sheer size and dynamism of WP which, compared with other static or slowly evolving resources, is constantly growing in size and in coverage regarding the emerging technologies.

Question answering (the automatic answering of a question posed in natural language) is an information retrieval approach that calls up not a collection of documents but a relevant sentence or fragment providing a specific answer. Authors like Kaisser (2008) include, in their usual Internet search procedure, structured WP information that helps establish a key factor: the type of entity sought as a response.

Wu *et al.* (2007), among others, have proposed, for the purpose of intelligent web searching, revitalizing the semantic web project (Berners-Lee *et al.*, 2001), until now impractical given the magnitude of the task of semantically tagging the entire web. These authors propose semantically tagging the WP hyperlink structure itself to convert it into a machine-readable structure of knowledge and could constitute the core for a real semantic web of the future.

1.3. Expanding and improving semantic knowledge bases

The WP structure facilitates the maintenance and expansion of knowledge structures such as the thesauri and ontologies that are essential for many tasks and applications. Exploitation of hierarchical links and redirect links enables synonyms and inclusion relations between concepts to be established and also enables new concepts not covered by other ontologies to be explored. Infoboxes and the categories structure enable WP content to be converted directly into relational databases. Moreover, opening paragraphs in WP articles provide glosses (descriptions of concepts), interlingual links make it possible to create equivalent knowledge bases in other languages and comparison of the WP structure with that of another ontology permits expansion of both.

Information extraction, aimed at building structured knowledge bases from unstructured text, is optimized by exploitation of the encyclopedic nature of WP and its uniform writing style.

A more ambitious goal is exploitation of the structure of WP links and categories in order to explicitly represent facts and rules and augment them with legacy systems that capture the overall semantic structure of language. This structure, which can be extracted from WP, is less costly and offers more coverage than manually developed knowledge bases such as WordNet and CYC. The three main approaches currently being investigated are extraction and tagging of WP categorical relations, extraction and organization of infobox information and enrichment of manually developed resources with information extracted from WP.

The main project in this regard is YAGO (Suchanek *et al.* 2007), a large taxonomy created by enriching concept structures in WordNet with categories and articles from WP. These authors are also enriching WordNet with semantic relations that were not previously in WordNet but were in WP, such as "born_in_year" and "born_in_place". The authors claim to have created a knowledge base containing 20 million facts attributed to two million entities and equipped, moreover, with a highly expressive descriptive logic that allows inferences to be drawn from the facts. Another project along the same lines is DBpedia (Auer *et al.* 2007), which has transformed WP infoboxes into a giant set of logical propositions (103 million).

1.4. Educational applications

Sawaki *et al.* (2007) developed an answer search alternative that could potentially be applied to teaching: the extraction of questions, answers and clues from biographical WP articles with which to automatically generate puzzles of the kind "Who is this person?". An interesting line of work in this regard is the ordering of clues according to difficulty so that the puzzle can be adapted accordingly.

Moré (2009) from the UOC exploits the hierarchical structure of WP categories to generate semantic concept fields (which can be multilingual) that can be used to prepare teaching and other academic materials. Moré *et al.* (2010) also use WP as an online teacher support resource via techniques that obtain answers based on students' messages as questions and using WP and other sources as a collection of documents in which to find answers.

1.5. Multilingualism

Given its multilingual nature and the interlingual connection between entries, WP is both a multilingual dictionary and an aligned corpus. It can therefore be potentially used for machine translation purposes or for interlingual information retrieval (retrieval of information in a language other than that in which the query is launched) through the multilingual creation and expansion of questions.

Adafre *et al.* (2006) have developed another interesting use of WP. Working in English and Dutch, they compare machine translations of WP articles with the corresponding article written by a human and create bilingual lexicons that are used to identify pairs of sentences in different languages with the same or similar meaning. Erdmann *et al.* (2008) exploit WP's structure in a sophisticated way by going beyond direct interlingual links to create bilingual dictionaries.

Below we describe research of our own that exploits multilingualism in WP – currently underexplored – for the purpose of creating knowledge bases.

2. WordNet

As mentioned earlier, WordNet (Fellbaum, 1998) is a lexical database of English vocabulary, in which open-category words (nouns, verbs, adjectives and adverbs) are organized as sets of synonyms called "synsets". Each synset represents a lexicalized concept that is connected with other synsets through semantic relations. The main WordNet relations are as follows: hyponymy (a type-of relation between a hyponym and a more generic hypernym), antonymy (a relation that implies a directly opposite meaning), meronymy (a part-of relation between a part and a whole) and troponymy (a relation between verbs that is similar to hyponymy in nouns).

For example, the WordNet 3.0 English synset, as identified by an offset and grammar category 02958343-n, has several variants: *car, auto, automobile, machine* and *motorcar*. Each synset is assigned a gloss or definition and examples of usage are frequently provided. For this synset, for example, the information provided is *a motor vehicle with four wheels; usually propelled by an internal combustion engine* along with the example *he needs a car to get to work.* This synset has 31 hyponyms, including 02701002-n (*ambulance*) and 03594945-n (*jeep, landrover*). It also has a hypernym, namely, 03791235-n (*motor vehicle, automotive vehicle*). An example meronym is 02685365-n (*airbag*).

To give an example of troponymy, we need to consider a verbal synset, for example, 01926311-v (*run; move fast by using one's feet, with one foot off the ground at any given time*). The troponymy relation is specified in the form of hyponyms and hypernyms (12 and 1, respectively, according to WordNet), for example, 01928579-v (*sprint*; *run very fast, usually for a short distance*) and 02055649-v (*travel rapidly, speed, hurry, zip*; *move very fast*).

Synonymy occurs between all the lexical forms (variants) of a specific synset; for example, *car, auto, automobile, machine* and *motorcar* are synonyms in that they are variants of the synset 02958343-n.

WordNet has become a standard resource for all kinds of NLP semantic research and applications. WordNet in English is free and can be downloaded from the Princeton University website (http://wordnet.princeton.edu). In the remainder of this article, we will refer to this Princeton WordNet as PWN. The current version is 3.0, released in December 2006. Table 1 shows the number of synsets available in versions 1.5, 1.6 and 3.0 of the PWN.

	Version 1.5	Version 1.6	Version 3.0
Total	76,705	99,642	118,695
Nouns	51,253	66,025	83,073
Verbs	8,847	12,127	13,845
Adjectives	13,460	17,915	18,156
Adverbs	3,145	3,375	3,621

Table 1. Number of synsets for different versions of the Princeton University WordNet

As can be observed, the number of synsets has increased with each new version. Such updates also require word nets for other languages to be updated to ensure comparability.

Not all word nets are built and edited under a free license. Bond (2012), who lists existing word nets for different languages and the associated licences, indicates that the languages for which free word nets are available are English, Finnish, Russian, Thai, Danish, Japanese, Catalan, Irish Gaelic, Hindi, French, Malay, Indonesian, Spanish, Arabic and Hebrew.

2.1. Word net building strategies

Below we review some of the strategies that have been used to build word nets for different languages (excluding that for English, considered the original word net). Vossen (1998) distinguishes between two general approaches to building word nets:

- **Merge:** An ontology is first generated for the language and interlingual relations are subsequently generated between this word net and the PWN.
- **Expand:** The variants associated with the PWN synsets are translated using various strategies. In this case it is not necessary to establish interlingual relations because the word net and the PWN are parallel.

Each strategy has associated advantages and disadvantages (Vossen, 1996). The expand strategy is technically easier and ensures greater compatibility between word nets in different languages. However, such word nets are heavily influenced by the PWN and propagate all the latter's errors and structural weaknesses. The merge strategy is more complex but allows for optimal use of available ontologies and thesauri.

2.2. The Catalan and Spanish word nets

The first word nets for Catalan (Benitez *et al.,* 1998) and Spanish (Atserias *et al.*, 1997) were based on applying an expand strategy to PWN 1.5 based on translating the variants corresponding to PWN synsets. This translation was performed using bilingual dictionaries, which described relations between English words and Catalan (or Spanish) words, although not between synsets and Catalan (or Spanish) words. To establish levels of confidence in assigning Catalan (or Spanish) variants to synsets: (i) relations between English words and synsets were assigned to monosemic and polysemic groups, and (ii) relations between English and Catalan (or Spanish) words were grouped according to the number of translations. This allowed synsets and Catalan (or Spanish) words to be directly related (thus obtaining variants in Catalan or Spanish) and a confidence level to be assigned to each relation. Catalan and Spanish word net versions 1.6 were generated by mapping synsets from version 1.5 to synsets in version 1.6.

The main difference between the Catalan and Spanish word nets is the licence used for distribution. The Catalan word net is distributed under a free license (GNU-GPL), whereas the Spanish word net is available under a proprietary license, although a small portion is distributed under a free licence along with the Freeling analyzer (the complete version is freely available for research).

3. Using Wikipedia to build word nets

Below we describe three WP-based applications that we developed to enrich Catalan and Spanish word nets version 3.0 with new variants or concepts: data explopitation in a knowledge base associated with WP called Babelnet; WP used as a bilingual dictionary; and WP used as a source of invariable proper names.

3.1. The Babelnet project

The aim of the Babelnet project (Navigli *et al.*, 2010) is to create a large-scale semantic network that relates word-net lexical knowledge to WP encyclopaedic knowledge. In Figure 1 we can see how Babelnet relates synset 02958343-n for the variants *motorcar, automobile* and *car* with the English WP entry for *car*.

Figure 1. Relation between synset 02958343-n and the corresponding entry in Wikipedia

5558	Motorcar	motorcar%1:06:00::	02958343n
11802	Automobile	automobile%1:06:00::	02958343n
45193	Car	car%1:06:00::	02958343n

Based on this approach, whereby we have a relation between a synset *s* and an entry w_{eng} in the English WP, we can use interlingual links to establish the same relation for the equivalent entries in other languages (w_{cat}, w_{spa}, etc).

Figure 2 shows the interlingual links for the WP entry for *car* extracted from an XML dump. With this information we can relate the synset 02958343-n with the Catalan variant *automòbil*.

Figure 2. Fragment showing interlingual links corresponding to an entry

```
[[ast:Automóvil]]
[[gn:Mba'yruguata]]
[[az:Avtomobil]]
[[bn:গাড়ী]]
[[zh-min-nan:Chū-tōng-chhia]]
[[be: Аўтамабіль]]
[[be-x-old::Аўтамабіль]]
[[bs:Automobil]]
[[br:Karr-tan]]
[[bg: Автомо́бил]]
[[ca:Automòbil]]
[[cs:Automobil]]
[[co:Vittura]]
[[cy:Car]]
[[da:Bil]]
[[pdc:Maschien]]
[[de:Automobil]]
[[nv:Chidí]]
[[et:Auto]]
[[el::Αυτοκίνητο]]
[[es:Automóvil]]
```

We can then find other variants for the original entry. WP contains a number of redirect pages that link up entries to a main entry to which they are equivalent. As can be seen in Figure 3, which shows the structure in XML of the Catalan entry for *cotxe* (car), this entry is redirected to *automóbil*. In other words, if we look for *cotxe*, the information displayed will be the entry for *automòbil*. This redirect system allows us to establish, for example, that *cotxe* is a valid Catalan variant for the 02958343-n synset.

Figure 3. Information on the Catalan Wikipedia entry for cotxe

```
<page>
  <title>Cotxe</title>
  <id>19827</id>
  <redirect />
  <revision>
    <id>124331</id>
    <timestamp>2004-10-14T07:13:01Z</timestamp>
    <contributor>
      <username>Arnadí</username>
      <id>281</id>
    </contributor>
    <comment>#REDIRECT [[Automòbil]]</comment>
    <text xml:space="preserve">#REDIRECT [[Automòbil]]</text>
  </revision>
 </page>
```

In practice, however, it is difficult to take full advantage of this feature as redirect pages have a broader use. For *automòbil* in Catalan, for example, we have the following synonyms: *automoció, cotxe, cotxes, elements aerodinàmics en l'automòbil.* This example indicates that the information cannot be automatically used, although it certainly could be useful but provided the proposals are subsequently manually reviewed.

In some cases, the language of interest may not have a corresponding entry in WP. To obtain variants in this case, Navigli *et al.* (2010) use Google Translate to translate a number of sentences in English containing variants from the synset *s*. These sentences are taken from two sources: SemCor and WP.

- **The SemCor corpus** (Miller et al., 1993). This corpus of English is semantically tagged (that is, most, if not all, nouns, verbs, adjectives and adverbs have a tag that indicates meaning). Tags for this corpus are the synsets of WordNet. Figure 4 shows an example sentence from this corpus with its tags (note that the real format of the corpus is different). The SemCor corpus can be downloaded from http://www.cse.unt.edu/~rada/downloads.html.

Figure 4. Example sentence from the SemCor corpus (format adapted for easier comprehension).

Then/00117620r he noticed/0215408v that the dry/02551380a wood/15098161n of the wheels/0457499n had swollen/00256507v.

- WP sentences containing links to the page *weng.* WP texts contain many hyperlinks to WP pages. Figure 5 shows a fragment of the Catalan entry for *automòbil* with hyperlinks marked in blue. These hyperlinks can also be considered as a kind of semantic tag, in that ambiguous words are linked to pages with the correct meaning. For example, the word *bateries* in Figure 5 could be considered ambiguous, but WP provides information about ambiguous words in a disambiguation page. Figure 6, for example, shows disambiguation information for *bateria*. The link for *bateries* in Figure 5 goes directly to the page giving information on electrical batteries and so confirms the meaning of this word in the text.

Figure 5. Fragment of the Catalan entry for automòbil with hyperlinks marked in blue.

L'**automòbil** (comunament **cotxe** o **votura** a la Catalunya Nord) és usualment un vehicle de quatre rodes destinat al transport de persones, amb capacitat entre dos i vuit seients. Es desplaça gràcies a un motor d'explosió a base d'una mescla de gasolina, gasoil i aire. En alguns països el combustible es fabrica a partir de determinades plantes en forma d'alcohol etílic. Recentment s'han començat a produir automòbils que funcionen amb motor elèctric, si bé l'autonomia d'aquests vehicles és encara limitada a causa del pes de les bateries. Les rodes davanteres dels automòbils poden moure's cap a ambdós costats per a fer girs i prendre les corbes.

Figure 6. Disambiguation information for the Catalan word bateria.

Bateria té els significats següents.

- *Electricitat:* Bateria elèctrica, conjunt d'acumuladors connectats en sèrie o en paral·lel
- *Exèrcit:* Bateria (unitat militar), agrupació tàctica i de tir elemental composta per un conjunt d'artillers i de canons
- *Música*: Bateria (instrument musical), conjunt d'instruments de percussió que és tocat per un sol instrumentista
- Eines: Conjunt d'estris de cuina. Vegeu la Categoria: Estris de cuina

Once these sentences have been machine-translated, the most common translation is identified and is included as a variant corresponding to the language of interest.

3.2. Using Wikipedia as a bilingual dictionary

Encyclopaedic bilingual dictionaries can be generated from WP, that is, dictionaries containing both general language words of encyclopedic interest and also the names of people, geographic locations, etc. Creating such dictionaries is simply a matter of following interlingual links. Just to provide readers with an idea of size, bilingual English-Catalan and English-Spanish dictionaries with 233,130 entries and 481,105 entries, respectively, can be generated from the English WP.

Such dictionaries can be used to directly assign Catalan and Spanish variants on the basis of English variants assigned to a specific synset. This assignation can only be determined with some degree of certainty in cases where a specific variant is assigned to a single synset, that is, when a word is monosemic. Table 2 summarizes monosemy and polysemy statistics for variants in WordNet 3.0 in English; note how the majority (82.32%) of the variants are, in fact, monosemic.

Number of meanings	Variants	%
1	123,228	82.32
2	15,577	10.41
3	5,027	3.36
4	2,199	1.47
5 or more	3,659	2.44
Total variants	149,691	100

Table 2. Monosemy and polysemy for variants in WordNet 3.0 in English

Of these monosemic variants, a certain proportion have corresponding interlingual links (from which the bilingual dictionary is determined) and so can be directly assigned a Catalan or Spanish variant (which should, however, be checked manually). Using this method, we generated 3,719 and 5,260 variants for Catalan and for Spanish, respectively.

3.3. Detecting invariable proper names

WordNet contains many variants corresponding to proper names; WordNet 3.0 in English, for example, contains over 40,000 variants written with a capital letter. Many of these variants are names of people and geographic locations that are written exactly the same in English, Spanish and Catalan; they are not listed, however, as entries in the Catalan or Spanish WPs and so do not have interlingual links to these languages.

We can use interlingual links to other languages to check whether, as well as in English, the variant in question is written the same in, say, five languages. In our initial experiments using this strategy, performed only with multi-word units, 2,858 and 2,928 variants were obtained for Catalan and Spanish, respectively, with an accuracy rate of above 87%.

4. Conclusions

We have described the current situation regarding the use made of WP for NLP-related tasks and, more specifically, for the creation of language resources. In regard to language resources, we have described three ways in which we use WP to automatically generate lexical knowledge bases for Catalan and Spanish from WordNet 3.0, by now a standard resource in semantic processing. The approaches have the advantage that they can be applied to any language with an associated WP. Although they will not enable the creation of complete word nets, they can be usefully combined with other manual or automated techniques.

References

ADAFRE, S.F.; DE RIJKE M. (2006). "Finding Similar Sentences across Multiple Languages in Wikipedia". In: *Proceedings of EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources*. pp. 62-69. Trento, Italy.

ATSERIAS, J.; CLIMENT, S.; FARRERES, X.; RIGAU, G.; RODRÍGUEZ, H. (1997) "Combining multiple methods for the automatic construction of multilingual wordnets". In: *Proceedings of RANLP'97*, pp. 143-149. Tsigov Chark, Bulgaria.

AUER, S.; BIZER, C.; LEHMAN, J.; KOBILAROV, G.; CYGANIAK, R.; IVES, Z. (2007). "DBpedia: A Nucleus for a Web of Open Data". In: Aberer *et al*. (Eds.): *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*. Busan, Korea. Lecture Notes in Computer Science 4825, Springer 2007.

BANERJEE S. (2007) "Boosting inductive transfer for text classification using Wikipedia". In: *Proceedings of the 6th International Conference on Machine Learning and Applications* (ICMLA), pp. 148-153. Cincinnati, Ohio, USA.

BENÍTEZ, L.; CERVELL, S.; ESCUDERO, G.; LÓPEZ, M.; RIGAU, G.; TAULÉ, M. (1998). "Methods and tools for building the Catalan wordnet." In: *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources and Evaluation*, Granada, Spain.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O.(2001). "The semantic web". *Scientific American.* Vol. 284, iss. 5, pp. 34-43. http://dx.doi.org/10.1038/scientificamerican0501-34

BOND, F.; KYONGHEE, P. (2012). "A Survey of WordNets and their Licenses". In: *Proceedings of the 6th International Global WordNet Conference*, pp. 64-71. Matsue, Japan.

DEERWESTER, S.; DUMAIS, S.T.; FURNAS, G.W.; LANDAUER, T.K.; HARSHMAN, R. (1990). "Indexing By Latent Semantic Analysis". *Journal of the American Society For Information Science*, iss. 41, pp. 391-407.
http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

ERDMANN, M.; NAKAYAMA, K.; HARA, T.; NISHIO, S. (2008). "An Approach for Extracting Bilingual Terminology from Wikipedia". In: *Proceedings of the International Conference on Database Systems for Advanced Applications* (DASFAA). New Delhi, India.
http://dx.doi.org/10.1007/978-3-540-78568-2_28

FELLBAUM C. (1998). "WordNet: An Electronic Lexical Database and some of its Applications." MIT Press.

GABRILOVICH, E.; MARKOVITCH, S. (2006). "Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge". In: *Proceedings of the 21st National Conference on Artificial Intelligence* (AAAI) pp. 1301-1306. Boston, USA.

GABRILOVICH, E.; MARKOVITCH, S. (2007). "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis". In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (IJCAI'07). pp. 1606-1611. Hyderabad, India.

GABRILOVICH, E.; MARKOVITCH, S. (2009). "Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research.* Iss. 34, pp. 443-498.

GREGOROWICZ, A.; KRAMER, M.A. (2006). "Mining a Large-Scale Term-Concept Network from Wikipedia". Technical Report 06-1028, Mitre.

GILES, J. (2005). "Internet encyclopaedias go head to head". *Nature*. Iss. 438, pp. 900-901.
http://dx.doi.org/10.1038/438900a

KAISSER, M. (2008). "The QuALiM Question Answering Demo: Supplementing Answers with Paragraphs drawn from Wikipedia". In: *Proceedings of ACL* (Demo Papers). pp. 32-35. Columbus, Ohio, USA.

LANGACKER, R. (1987). *Foundations of Cognitive Grammar, Volume I.* Stanford CA: Stanford University Press.

LENAT, D.; GUHA, R.V. (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project.* Addison-Wesley.

MEDELYAN, O.; WITTEN, I.H.; MILNE, D. (2008). "Topic Indexing with Wikipedia". *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy.* AAAI Press, Chicago, USA.

MEDELYAN, O.; MILNE, D.; LEGG, C.; WITTEN, I.H. (2009). "Mining Meaning from Wikipedia" *International Journal of Human-Computer Studies* 67:716-754.
http://dx.doi.org/10.1016/j.ijhcs.2009.05.004

MILLER, G.A. [et al.] (1993). "A semantic concordance". In: *Proceedings of the Workshop on Human Language Technology*. pp. 303–308. Stroudsburg, Pennsylvania, USA.
http://dx.doi.org/10.3115/1075671.1075742

MORÉ, J. (2009). "Creació automàtica de diccionaris multilingües especialitzats en noves àrees temàtiques" [online article]. Digithum. No. 11. UOC. [Accessed: 27/02/2012].
<http://www.uoc.edu/ojs/index.php/digithum/article/view/n11_more/n11_more>

MORÉ, J.; CLIMENT, S.; COLL-FLORIT, M.; RIVERA, J. (2010). "A Question-Answering Environment for eLearning Tutors". *International Journal of the Computer, the Internet and Managment* (IJCIM) ISSN 0858-7027. Vol 18 No. SP1. pp. 3.1-3.6. <http://lpg.uoc.edu/files/QA-Tutors-More-Climent-Coll-Rivera.pdf>

NAVIGLI, R.; PONZETTO, S.P. (2010). "BabelNet: building a very large multilingual semantic network". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL'10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 216–225. [Online. Accessed: 11 March 2011]. <http://portal.acm.org/citation.cfm?id=1858681.1858704>

PONZETTO, S.P.; STRUBE, M. (2006). "Exploiting semantic role labelling, WordNet and Wikipedia for coreference resolution". In: *Proceedings of HLT-NAACL'06.* pp. 192-199. New York, USA.

SAWAKI, M.; MINAMI, M.Y.; HIGASHINAKA, R.; DOHSAKA, K.; YAMADA, T.; MATSUBAYASHI, T.; ISOZAKI, H.; MAEDA, E. (2007). "Quizmaster Mushrooms: 'Who is this' Quiz Dialogue System". International Conference on Multimodal Interaction ICMI (Demonstration). Nagoya, Japan.

SUCHANEK, F.M.; KASNECI, G.; WEIKUM, G. (2007). "Yago - A Core of Semantic Knowledge". In: *Proceedings of the 16th International World Wide Web Conference* (WWW 2007). NewYork: ACM Press

THOMAS, C.S.; AMIT, P. (2007). "Semantic Convergence of Wikipedia Arrticles". In: *Proceedings of the International Conference on Web Intelligence* (IEEE/WIC/ACM WI'07). Hong Kong.

VIVALDI, J.; RODRIGUEZ, H. (2010). "Finding Domain Terms using Wikipedia". In: *Proceedings of the 7th LREC International Conference.* pp. 386-393. Malta.

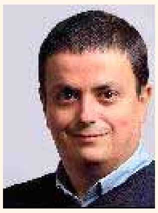
VOSSEN, P. (1996). "Right or Wrong. Combining lexical resources in the EuroWordNet project". In: *Proceedings of Euralex-96.* pp. 715–728. Gothenburg, Sweden.

VOSSEN, P. (1998). "Introduction to Eurowordnet." *Computers and the Humanities.* Vol. 32, iss, 2, pp. 73–89.
http://dx.doi.org/10.1023/A:1001175424222

WU, F.; WELD, D. (2007). "Autonomously semantifying Wikipedia". In: *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (CIKM'07). pp. 41–50. Lisbon, Portugal.

RECOMMENDED CITATION

OLIVER, Antoni; CLIMENT, Salvador (2012). "Using Wikipedia to develop language resources: WordNet 3.0 in Catalan and Spanish". In: Eduard AIBAR and Mayo FUSTER (coords.). "Academic research into Wikipedia" [online dossier]. *Digithum* No. 14, pp. 68-76. UOC. [Accessed: dd/mm/yy].
<http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-oliver-climent/n14-oliver-climent-en>
http://dx.doi.org/10.7238/d.v0i14.1474
ISSN 1575-2275



Antoni Oliver
Lecturer of the Department of Arts and Humanities
and director of the postgraduate course on Translation and Technologies (UOC)
aoliverg@uoc.edu

Antoni Oliver is lecturer of the Department of Arts and Humanities and director of the postgraduate course on Translation and Technologies at the Open University of Catalonia (UOC). His research interest is computational linguistics, especially in relation to machine translation.

Estudis d'Arts i Humanitats
Universitat Oberta de Catalunya
Avinguda Tibidabo 39-43
08035 Barcelona

Salvador Climent Roca
Lecturer of the Department of Arts and Humanities (UOC)
scliment@uoc.edu

Salvador Climent is lecturer of the Department of Arts and Humanities at the Open University of Catalonia (UOC). He coordinates the Catalan language and literature degree course and teaches general and cognitive linguistic subjects. He conducts research into computational linguistics and cognitive linguistics.

Estudis d'Arts i Humanitats
Universitat Oberta de Catalunya
Avinguda Tibidabo 39-43
08035 Barcelona

UOC
Universitat Oberta de Catalunya