WordNet Applications

Jorge Morato[1], Miguel Ángel Marzal[2], Juan Lloréns[1], and José Moreiro[2]

[1] Dept. Computer Science, Universidad Carlos III, Madrid, Spain
Email: jorge@ie.inf.uc3m.es, llorens@ie.inf.uc3m.es
[2] Dept. Library Science, Universidad Carlos III, Madrid, Spain
Email: mmarzal@bib.uc3m.es, jamore@bib.uc3m.es

Abstract. This paper describes WordNet design and development, discussing its origins, the objectives it initially intended to reach and the subsequent use to which it has been put, the factor that has determined its structure and success. The emphasis in this description of the product is on its main applications, given the instrumental nature of WordNet, and on the improvements and upgrades of the tool itself, along with its use in natural language processing systems. The purpose of the paper is to identify the most significant recent trends with respect to this product, to provide a full and useful overview of WordNet for researchers working in the field of information retrieval. The existing literature is reviewed and present applications are classified to concur with the areas discussed at the First International WordNet Congress.

1 Introduction

WordNet, one of a series of manually compiled electronic dictionaries, is restricted to no specific domain and covers most English nouns, adjectives, verbs and adverbs. Although there are similar products, such as Roget's International Thesaurus, or CYC, Cycorp: Makers of the Cyc Knowledge Server for artificial intelligence-based Common Sense. CYC contains 100,000 concepts and thousands of relations. Each concept is assigned to certain terms related by prepositional logic. The present paper analyses the reasons for WordNet's success and, in particular, the main applications of the tool over the last ten years.

2 Wordnet Development

The origin of this tool is to build a lexical-conceptual model and database, consisting of both lexical units and the relations between such units, structured into a relational semantic network.

Originally intending to create a product that could combine the advantages of electronic dictionaries and on-line thesauri, an expert team of linguists and psycholinguists headed by G. A. Miller began research at Princeton University's Cognitive Science Laboratory in 1985 that would culminate in the appearance of WordNet in 1993.

WordNet offers researchers, many of which were not initially envisaged by the authors, along with its cost-free use and well-documented open code. The result has been the appearance of applications in different fields of research, making it an ideal tool for disambiguation of meaning, semantic tagging and information retrieval. Therefore, although four members manage, maintain and develop WordNet many other teams collaborate in driving implementation of the product, as attested by two facts:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 270–278.

1. The speedy pace of release of new versions of WordNet.
2. Organised world-wide promotion of WordNet, through the creation of the *Global WordNet Association*, which, in conjunction with CIIL Mysore, IIT Bombay and IIIT Hyderabad, held the 1st International WordNet Conference in 2002. Primarily technical, the conference was structured under six areas of interest: Linguistics, WordNet architecture, WordNet as a lexical resource and component of NLP, Tools and Methods for WordNet Developme, nt., Standardisation., Applications (information extraction and retrieval, document structuring and categorisation, language teaching).

These six topics are still present in the 2nd International Conference of the Global WordNet Association (GWC 2004) held at Masaryk University, Brno.

3 Applications

The success of WordNet, as mentioned, is largely due to its accessibility, quality and potential in terms of NLP. Figure 1 below shows the results of a search run on the bibliographic database LISA, INSPEC, IEEE and ResearchIndex and on the Universidad Carlos III's OPAC. The documents were publised from 1994 till 2003. This search, while not necessarily exhaustive in respect of WordNet research, does nonetheless show how such research effort is distributed. It will be observed that the major use of this tool has been in the area of conceptual disambiguation.

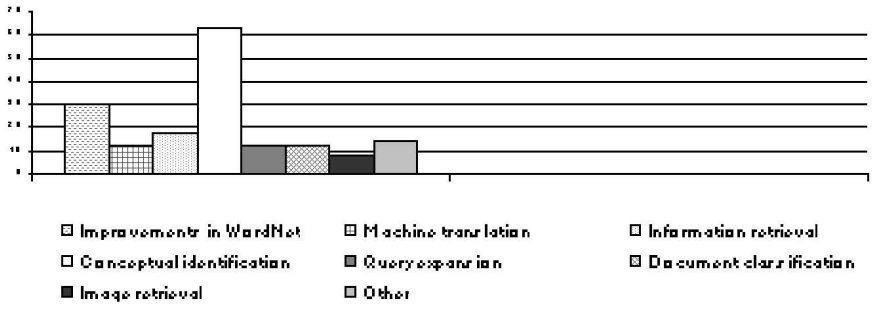

Fig. 1. WordNet Applications

3.1 Improvements in WordNet

The data record of publications dealing with WordNet shows that there has been a tendency to improve the product in a number of respects. The objective is to make WordNet much more effective and relevant than any existing on-line dictionary by incorporating greater semantic wealth and taking a more contextual approach. Several possibilities have been explored to achieve this:

Studies geared to improving software There is a clear prevalence, in terms of volume, of papers geared to expanding and enriching the WordNet structure. One of such endeavours is VerbNet [1], designed to make up for shortcomings in the associative relations between verbs; another is the Lingua::WordNet interface [2], which furnishes an editable presentation of WordNet, with meronym categories never before implemented. Finally, substantial efforts have been made to standardise, integrate and distribute the tool.

Multilingual WordNet One of the most relevant endeavours has been the development of EuroWordNet, a project based on WordNet structure whose ultimate purpose is to develop multilingual databases with wordnets for several European languages. Each wordnet adopts an autonomous lexicalisation structure and all are interconnected through an interlinguistic index, for which relations have been added and modified and new levels identified in WordNet. For a multilingual description of EuroWordNet see [3,4].

This paper poses the possibility of automatically transferring a list of English verbs, classified by their syntactic characteristics, to WordNet synsets.

3.2 Improvements in Natural Language Processing Systems

Such improvements are found in a substantially larger number of papers on WordNet, regarded here to be a tool well suited to a series of applications such as discussed below:

Information retrieval and extraction These operations are closely related to organisation and representation of knowledge on the Internet. One of the lines of research pursued is the application of artificial intelligence to information retrieval, stressing the local components and inferential process of human reasoning in the design of automatic information retrieval systems. The method for incorporating logic and inference focused on WordNet shortly after it appeared [5]. WordNet has been used as a comprehensive semantic lexicon in a module for full text message retrieval in a communication aid, in which queries are expanded through keyword design [6]. WordNet has, then, started to be used as a linguistic knowledge tool to represent and interpret the meaning of, and provide the user with efficient and integrated access to, information; integration, indeed, has become an increasingly necessary feature with the development of multiple database access systems and one in which WordNet's identification and interpretation of semantic equivalents is extraordinarily useful [7].

Mandala [8] proposed the use of WordNet as a tool for the automatic construction of thesauri, based either on co-occurrence determined by automatic statistical identification of semantic relations, or on the predicate-argument association, in which the most significant words of an environment (predicate) and those with which they relate are identified to construct the argument. In another vein, Moldovan [9] opted to use WordNet in the development of a natural language interface to optimise the precision of Internet search engines by expanding queries.

Concept identification in natural language This operation is designed to detect the terms requested, not only for extraction, but to suit them to the full semantic richness and complexity of a given information need. WordNet applications have followed a dual course in such applications:

1. **Disambiguation** i.e., precision and relevance in response to a query via resolution of semantic inconsistencies. Moldovan [9] described schematically the semantic disambiguation as follows:
   (1) All the noun–verb pairs in the sentence are selected.
   (2) The most likely meaning of the term is chosen (subprocess that Moldovan calls terminological disambiguation). Internet is used with this goal.
   (3) Drawing from the most frequently appearing concepts (step 2), all the nouns are selected in the "glossaries" of each verb and its hierarchical subordinates.
   (4) Drawing from the most frequently appearing concepts, all the nouns are selected in the "glossaries" of each noun and its hierarchical subordinates.
   (5) A formula is applied to calculate the concepts common to the nouns in points 3 and 4.

   Disambiguation is unquestionably the most abundant and varied WordNet application. Indeed, there is a wide range of possibilities.
   WordNet has served as a support for the development of tools to enhance the efficiency of Internet resource searches. One example is the IWA/H project for building an ontological framework able to disambiguate search criteria via mixed knowledge representation technique systems (ARPA KRSL); others include tools such as Oingo and SimpliFind, two Internet products that avoid ambiguity in natural language searches by using the WordNet lexicon, duly expanded by creating millions of word associations to refine the search process.
   The use of WordNet for improving search engines is interesting the IWA/H project was based on the MORE technique developed by the RBSE project for more efficient retrieval of Internet resources, as discussed by Eichmann [10].
   WordNet has, naturally, been used for disambiguation in traditional models to enhance information retrieval efficiency: for the development of a classifier, implemented with WordNet, able to combine a neurone-like network to process subject contexts and a network to process local context; for the exploitation of a Bayesian network able to establish lexical relations with WordNet as a source of knowledge, integrating symbolic and statistical information [11]; for the development of a statistical classifier, implemented with WordNet lexical relations, able to identify the meaning of words, combining the context with local signs [12]; and as support for the development of a computational similarity model to add on-line semantic representation to the statistical corpus. WordNet has, therefore, proved its worth as an ideal methodological element to disambiguate the meaning of words in information extraction systems [13]. As a result, projects have been launched to disambiguate nouns in English language texts using specification marks deriving from WordNet taxonomies as a knowledge base, as well as to reduce polysemy in verbs, classified by their meanings via predicate associations, with a view to optimising information retrieval. Methods for nouns [14] and verbs [1,4] has already been analysed.
   At the same time, new disambiguation models have been tested in conjunction with WordNet by: generating ontological databases with a systematic classification of multiple meanings derived from WordNet [15]; or generating broad corpora to signify words on the grounds of WordNet synonymies or definitions in the wording of queries [16]. One result has been the appearance of GINGER II, an extensive dictionary semantically tagged using 45 WordNet categories and an algorithm for interpreting

semantic text by determining verb senses, identifying thematic roles and joining prepositional phrases [17]. More recently R. Mihalcea and D. Moldovan presented AutoASC, which automatically generates sense tagged corpora that prove to be very effective for disambiguation in information retrieval; this product incorporates the latest WordNet gloss definitions [18].

2. **Semantic distance** Three concepts recur in WordNet literature that entail a certain amount of ambiguity: terminological distance, semantic distance and conceptual distance. The terms semantic distance and conceptual distance are found to be used in several studies to pursue the same objective and deploy the same methodology for resolving the issue at hand. Terminological distance, by contrast, often appears to refer to the suitability of the word selected to express a given concept.
   Semantic distance is understood to mean the contextual factor of precision in meaning. In his particularly relevant papers, Resnik [19] computes class similarity, defining class to be the nouns of a synset plus the nouns in all the subordinate synsets. Although the concept of semantic similarity between classes was proposed by Resnick [19]. WordNet was quickly enlisted to build and operate with FEDDICT, a prototype on-line dictionary to develop an information retrieval technique based on the measurement of the conceptual distance between words, in which WordNet semantic relations proved to be highly useful [20]. A very interesting sequel to this endeavour was provided by McTavish [21] who used WordNet semantic domains to establish categories that could be used to analyse conceptual semantic distances in terms of social environments to better organise terms for retrieval.
   Computational linguistics is, however, the area that has placed the greatest emphasis on *relations* and *semantic distances* between lexemes, the measures of which were classified by A. Budanitsky [22]. This paper highlights the measures that use WordNet as a resource and for implementation of functions, in particular: Hist-St. And Leacock–Chodorow, in which similarity, albeit in the IS-A link only, rests on the shortest path between two synsets; and Resnik, Jiang, Conrath and Lin, for all of whom *information content* is a determining factor of similarity in their measures of distance.

Query expansion In 1994 Smeaton [23] proposed an expansion system based on calculating the tf-idf for the query terms and adding to it half the tf-idf for the WordNet synonyms for these terms. Gonzalo [24] later reported the benefits of applying WordNet to queries, using it as a WSD (Word Sense Disambiguator) able to enhance the search process by including semantically related terms and thus retrieve texts in which the query terms do not specifically appear.

Document structuring and categorisation Intellective efforts and operations in this area are geared to a new organisation and representation of knowledge. In this case the focus is on the aspects of the tool suited to document categorisation: extraction of semantic traits by grammatical categorisation of WordNet nouns, verbs and adjectives [25]; and categorisation of the relevance of the data in INFOS by predicting user interest on the basis of a hybrid model using keywords and WordNet conceptual representation of knowledge [26].

Further research along these lines came in the form of a computational method presented by S. M. Harabagiu [27] for recognising cohesive and coherent structures in texts, drawing on

WordNet lexical-semantic information, whose objective is to build designs for the association between sentences and coherence relations as well as to find lexical characteristics in coherence categories. WordNet became an ancillary tool for semantic ontology design geared to high quality information extraction from the web, and has prompted new endeavours such as the WebOntEx (Web Ontology Extraction) prototype developed by Keng Woei Tan [28] which is designed to create ontologies for the semantic description of data in the web.

Judith Klavans [29] devised an algorithm for automatically determining the genre of a paper on the grounds of the WordNet verb categories used. With their WN-Verber, they determined that some verbal synsets and their highest subordinates are less frequent in certain document typologies.

Audio and video retrieval This is a challenge in need of increasingly urgent attention in view of the burgeoning development of hypermedia and non-text information. The MultiMediaMiner [30], is a prototype to extract multimedia information and knowledge from the web that uses WordNet to generate conceptual hierarchies for interactive information retrieval and build multi-dimensional cubes for multi-media data. Finally, WordNet has been used in query expansion to index radio news programme transcriptions effected by a prototype designed to retrieve information from radio broadcasts [31].

Other WordNet applications

Parameterisable information systems While anecdotal, the J. Chai [32] proposal to create an information system (called Meaning Extraction System) that can be configured in terms of a specific user profile is appealing. The user chooses from a series of texts (training collection) the ones that appear to be of greatest interest. WordNet identifies the hierarchical (IS-A) synsets related to the terminology of the documents selected. This process generates rules that enable the system to identify, *a priori*, the documents that the user will find to be of interest.

Language teaching and translation applications As discussed in point 3.1.2, applications have been devised and tested to improve the composition of texts drafted by non-native English writers. However, yet another line of research has been addressed in international conferences on WordNet, namely, foreign language teaching. One example is the article by X. Hu and A. Graesser, which proposes using the WordNet vocabulary to evaluate pupils' command of a given language [33].

As a translation aid based on the application of semantic distance algorithms, WordNet has also been used to develop a potential error corrector for the positioning of words [34]. One very intuitive formula consists of using *conceptual interlingua* representation of texts and queries such as used in the CINDOR system, which accommodates WordNet-supported inter-linguistic combinations, obviating the need for an expert translation for retrieval. The CINDOR system was presented and tested at TREC-7 and seems to be useful for cross- or combined linguistic retrieval [35].

4 Trends

Trends are difficult to ascertain and evaluate in view of the clearly instrumental and application-based dimension that underlies WordNet's success. Nonetheless, a comparative analysis of the most recent publications provides some insight into a number of trends in WordNet use:

1. Development of interlinguistic indices for multilingual conceptual equivalence, without translation. Subsidiarily, this endeavour has also been geared to perfecting integrated access to information, driven by the rapid development of multiple database access systems.
2. Use as an ideal tool to optimise the retrieval capacity of existing systems: natural language interfaces for search engines; automatic generation of tools for semantic disambiguation of concepts (corpora, dictionaries, directories, thesauri) and the creation of knowledge summaries from expanded queries.
3. Support for the design of grammatical categorisations designed to classify information by aspects and traits, but in particular to design and classify semantic ontologies that organise web data – semantically, to be sure.
4. Basis for the development of audio-visual and multi-media information retrieval systems.
5. In the last 3 years ontologies construction have been one of the most dynamic areas and its applications to the semantic web [36].

5 Conclusions

Although WordNet applications are growing steadily and research may be expected to increase in the coming years as new versions are released, the tool has certain shortcomings that should be addressed in future studies.

Limitations and Problems are:

1. Although WordNet is an electronic resource, it was, after all, designed for manual consultation and not for automatic processing of natural language texts; as a result, no particular emphasis was placed on enabling the system to automatically differentiate between the various concepts involved.
2. Another problem is its multidisciplinarity, which prompts flawed operation in many NLP systems, due to which processing is usually conducted with sublanguages or specific records.
3. Classification was performed manually, which means that the reasons and depth of classification may not be consistent.
4. While the synset simplification affords obvious advantages, in the longer term it leads to shortcomings. These are particularly acute in semantic proximity calculations and may create insuperable situations whenever the context of the discourse in which the relation appears is not contained in the synset information.
5. The overabundance of nuance in the concepts calls, in nearly any NLP application, for prior calculation of the frequency of the concept in a given domain. Such calculation is one of the sources of system error, especially where WordNet glosses – extracted, as noted above, from the Brown Corpus – are used, due to the uneven coverage afforded to the different domains.

References

1. Palmer, M.: Consistent criteria for sense distinctions. Computers and the Humanities, 34 (1–2) (2000) 217–222.
2. Brian, Dan: Lingua::WordNet. The Perl Journal (2000) `http://www.brians.org/wordnet/article/`.
3. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities, 32(2–3) (1998) 73–89.
4. Green, Rebecca, Pearl, L., Dorr, B. J., and Resnik, P.: Mapping lexical entries in verbs database to WordNet senses. Proceedings of the 39þ Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse, France, July 9–11 (2001).
5. Nie, J. Y., and Brisebois, M.: An inferential approach to information retrieval and its implementation using a manual thesaurus. Artificial Intelligence Review, 10 (5–6) (1996) 409–439.
6. Van de Riet, R., Burg, H., and Dehne, F.: Linguistic instruments in information system design. FOIS. Proceedings of the 1st International Conference. Amsterdam: IOS Press (1998).
7. Jeong-Oog-Lee & Doo-Kwon-Baik: Semantic integration of databases using linguistic knowledge. Proceedings Advanced Topics in Artificial Intelligence. Berlin: Springer-Verlag, (1999).
8. Mandala, Rila, Tokunaga, T., Tanaka, Hozumi, O., Akitoshi, Satoh, K.: Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri. TREC-7 (1998) 414–419.
9. Moldovan, D. I. and Mihalcea, R.: Using WordNet and lexical operators to improve Internet searchers. IEEE Internet Computing, 4 (1) (2000) 34–43.
10. Eichmann, David: Balancing the Need for Knowledge and Nimbleness in Transportable Agents. Position Paper for the Dartmouth Workshop on Transportable Agents. (1996). URL: `http://mingo.info-science.uiowa.edu/eichmann/DWTA96.html`, last hit on 2/3/02.
11. Wiebe, Janyce, O'Hara, Tom, and Bruce, Rebecca: Constructing Bayesian Networks from WordNet for Word-Sense Disambiguation: Representational and Processing Issues. Use of {W}ord{N}et in Natural Language Processing Systems: Proceedings of the Conference Association for Computational Linguistics, Somerset, New Jersey (1998). 23–30.
12. Towell, G. and Voorhees, E. M.: Disambiguating highly ambiguous words. Computational Linguistics, 24 (1) (1998) 125–145.
13. Chai, Joyce Y. and Biermann, Alan W.: A WordNet based rule generalization engine for meaning extraction, to appear at Tenth International Symposium On Methodologies For Intelligent Systems (1997).
14. Montoyo, A. and Palomar, M.: Word sense disambiguation with specification marks in unrestricted texts. Proceedings 11th International Workshop on Database and Expert Systems Applications. Los Alamitos (Ca): IEEE Press (2000) 103–107.
15. Buitelaar, P.: CORELEX: an ontology of systematic polysemous class. Proceedings FOIS'98. Amsterdam: IOS Press. (1998) 221–235.
16. Mihalcea, R. and Moldovan, D. I.: Automatic acquisition of sense tagged corpora. Proceedings of the 12th International Florida AI Research Society Conference. Menlo Park(Ca): AAAI Press (1999) 293–297 and 16th: 461–466.
17. Dini, L., Tomasso, V., and Segond, F.: GINGER II: an example-driven word sense disambiguator. Computers and the Humanities, 34 (1–2) (2000). 121–126.
18. Mihalcea, R. and Moldovan, D. I.: AutoASC, a system for automatic acquisition of sense tagged corpora. International Journal of Pattern Recognition and Artificial Intelligence, 14 (1) (2000) 3–17.
19. Resnick, P: Selection and Information: A class-based approach to lexical relationships. PhD dissertation. University of Pennsylvania (1993).
20. Richardson, R., Smeaton, A. F., and Murphy, J: Using WordNet for conceptual distance measurement. Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group. London: Taylor Graham. (1996) 100–123.

21. McTavish, D. G., Litkowski, K. C. and Schrader, S: A computer content analysis approach to measuring social distance in residential organizations for older people. Social Science Computer Review, 15 (2) (1997) 170–180.
22. Budanitsky, A. and Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. (2001). URL: `http://www.cs.toronto.edu/pub/gh/`.
23. Smeaton, Alan F., Kelledy, Fergus, and O'Donell, Ruari: TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish. Proceedings of TREC-4. Gaithersburg (USA): D. Harman (Ed.) (1994).
24. Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J.: Indexing with WordNet synsets can improve text retrieval. Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP. Montreal (Canada) (1998) 38–44.
25. Scheler, G.: Extracting semantic features from unrestricted text. WCNN'96. Mahwah (NJ): L. Erlbaum. (1996).
26. Mock, K. J. and Vemuri, V. R: Information filtering via hill climbing, WordNet and index patterns. Information Processing and Management, 33 (5). (1997) 633–644.
27. Harabagiu, S. M.: WordNet-based inference of contextual cohesion and coherence. Proceedings of the 11th International Florida Artificial Intelligence Research Symposium Conference. Menlo Park (Ca): AAAI Press (1998) 265–269.
28. Keng Woei Tan, Hyoil-Han and Elmasri, R.: Web data cleansing and preparation of ontology extraction using WordNet. Proc. 1st International Conference on Web Information Systems Engineering. Los Alamitos (Ca): IEEE Computational Society, 2. (2000) 11–18.
29. Klavans, Judith and Kan, Min-Yen: Role of verbs in document analysis. Proceedings of the Conference, COLING-ACL. Canada: Université de Montreal. (1998).
30. Zaiane, O. R., Hagen, E., and Han, J.: Word taxonomy for online visual asset management and mining. Application of Natural Language to Information Systems. Proc. 4th Internat. Conference NLDB'99. Vienna: Osterreichische Comput. Gesseleschaft (1999) 271–275.
31. Federico, M.: A system for the retrieval of Italian broadcast news. Speech Communication, 32 (1–2) (2000). 37–47.
32. Chai, Joyce Y. and Biermann, Alan W.: The use of word sense disambiguation in an information extraction system. Proceedings 16th National Conference on Artificial Intelligence. Menlo Park (Ca): AAAI Press (1999) 850–855.
33. Hu, X & Graesser, A.: Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in tutorial dialogue. Proceedings of ICCE'98, 2. Beijing: China Higher Education Press (1998) 337–341.
34. Shei, C. C. and Pain, H.: An ESL writer's collocational aid. Computer Assisted Language Learning, 13 (2) (2000) 167–182.
35. Diekema, A., Oroumchian, F., Sheridan, P., and Liddy, E. D.: TREC-7 evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. Gaitherburg (USA): TREC-7 National Institute of Standards & Technology (1999) 169–180.
36. Khan, L, Luo, F: Ontology construction for information selection. Proceedings of the 14 IEEE ICTAI 02 (2002).