Knowledge Representation and WordNets

Alexandra Gabriela Tudorache –
alexandra.tudorache@gmail.com
PhD Student Academy of Economic Studies of Bucharest & University Tor Vergata of Rome

1. Knowledge Representation

Knowledge itself is a representation of “real facts”.[1]

Knowledge is a logical model that presents facts from “the real world” witch can be expressed in a formal language. Representation means the construction of a model of some part of reality.

Knowledge representation is contingent to both cognitive science and artificial intelligence. In cognitive science it expresses the way people store and process the information. In the AI field the goal is to store knowledge in such way that permits intelligent programs to represent information as nearly as possible to human intelligence.

Knowledge Representation is referred to the formal representation of knowledge intended to be processed and stored by computers and to draw conclusions from this knowledge.[2]

Examples of applications are expert systems, machine translation systems, computer-aided maintenance systems and information retrieval systems (including database front-ends).

1.1 Short History of Knowledge Representation

The evolution of computer science can be viewed as an evolution of data-representations towards knowledge representation.

In the 1970s and early 1980s major KR branches were developed like: heuristic question-answering, neural networks, theorem proving, and expert systems. Other important application areas were medical diagnosis (e.g., Mycin) and games (chess).

In the 1980s formal knowledge representation languages and systems have started to be developed. Some important projects like “Cyc” tried to encode wide bodies of general knowledge. "Cyc" project used a large encyclopedia encoding the information a reader would require and not the information itself. The Cyc project is still managed by Cycorp, Inc. a great part of the data being now freely available.

Beginning with projects like “Cyc” much and much larger linguistic resources were built and the KR became more feasible. This is the first step towards the Semantic Networks development.

Semantic networks can be used to represent knowledge. Each node of the network represents a concept and arcs are used to define relations between the concepts. In this area some important projects are: WordNet and MultiNet (Multilayered Extended Semantic Networks).

Since then several KR oriented programming languages have been developed: Prolog (1972), KL-ONE (1980s).

To represent the structure of electronic documents languages were being developed (e.g. SGML and later XML). These languages made possible information retrieval and data mining.

In the semantic Web XML-based languages like RDF, Topic Maps, OWL and others can be used to make Knowledge Representation information available to Web systems.

1.2 Language and notation

Some people think that the best solution is to represent knowledge like it is represented in the human mind so far the only known working intelligence. Another way to represent knowledge is in the form of human language. To accomplish this various languages and notations have been developed. Usually are based on logic and mathematics, and have easily parsed grammars to help machine processing. These languages are usually included in the broad area of Ontologies.

Examples of notations:

- DATR is an example for representing lexical knowledge
- RDF is a simple notation for representing relationships between and among objects

Examples of artificial intelligence languages used primarily for knowledge representation include:

- CycL
- IKL
- KIF

- Loom
- OWL
- KM : the Knowledge Machine (frame-based language used for knowledge representation work)

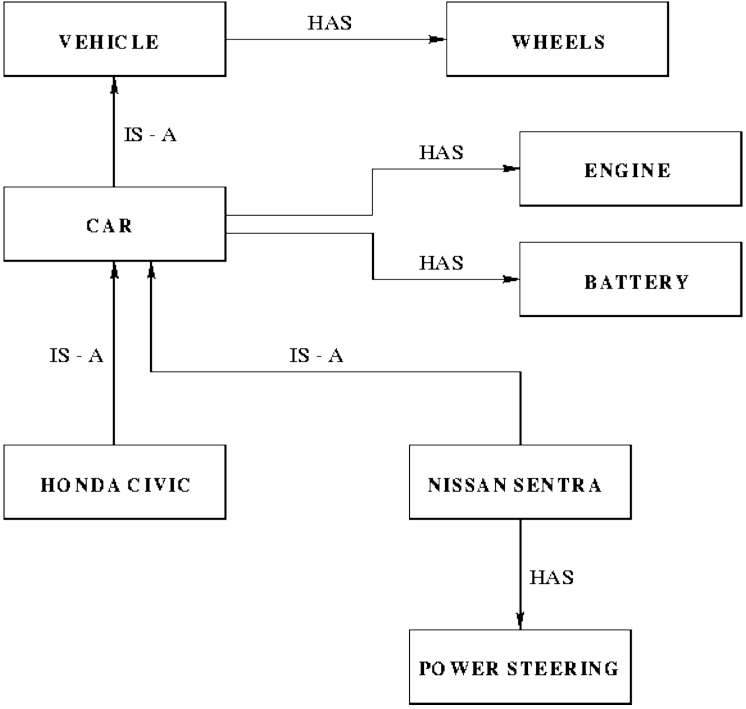
2. Semantic networks

Semantic networks are based on over fifty years of research in artificial intelligence, cognitive psychology, memetics and learning theory.

"Semantic Nets" were created in 1956 by Richard H. Richens (Cambridge Language Research Unit) as an "interlingua" for computerized translation of natural languages. In the early 1960s Robert F. Simmons took up their development at System Development Corporation, Santa Monica, California and later were brought up in the work of *M. Ross Quillian in 1966*.

In his Ph.D. thesis[3], Ross Quillian described a system for allowing the meaning of words to be modeled on a computer such that computational use of these meanings would be possible. This became the basis for the idea of a semantic network. Since then, several decades of research have refined the idea to its fullest modern expression.

A semantic network is a way of representing relationships between concepts. Each concept is expressed by a word or set of words. As an example let’s look at the semantic network representing a car. The concepts are the car itself, the parts of the car (engine, battery) and also some instances of car like Honda Civic, Nissan Sentra. Also the semantic Net encomprises taxonomic information in relations like is *type-of* (hyponymous relationship).



Example of a semantic network regarding cars

More complex semantic networks may include a variety of types of relationship such as hardness, temperature, made-of, texture and color. One of the largest existing semantic networks is WordNet[4], a lexical database for the English language.

There are also elaborate types of semantic networks connected with corresponding sets of software tools used for lexical knowledge engineering, like the Semantic Network Processing System (SNePS) of Stuart C. Shapiro or the MultiNet paradigm of Hermann Helbig (MultiNet is an acronym for "Multilayered Extended Semantic Network"). The latter is especially suited for the semantic representation of natural language expressions and used in several NLP applications.

In the 1960s to 1980s the idea of a semantic link was developed within hypertext systems as the most basic unit, or edge, in a semantic network. These ideas were extremely influential, and there have been many attempts to add typed link semantics to HTML and XML. This was the dawn of the Semantic web idea.

In computer science a semantic network is represented as a directed graph consisting of nodes (also termed points or vertices) which represent concepts and edges (also termed lines or arcs) which represent semantic relations between the concepts. [5]

Some important semantic relations:

- Hyponymy (or *troponymy*) (A is subordinate of B; A is kind of B)
- Hypernymy (A is superordinate of B, the inverse function of Hyponymy)
- Meronymy (A is part of B, i.e. B has A as a part of itself)
- Holonymy (B is part of A, i.e. A has B as a part of itself, is the inverse function of Meronymy)
- Synonymy (A denotes the same as B)
- Antonymy (A denotes the opposite of B)

3. WordNet

WordNet is a semantic lexicon for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets.

Nowadays WordNet is one of the most employed lexical resources as support for English Natural Language Processing (NLP).

WordNet it is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas. Its design is inspired by current psycholinguistic and computational theories of human lexical memory. English nouns, verbs, adjectives, and adverbs are organized into synonym sets called synsets. Each represents one underlying lexicalized concept.

Two kinds of relations that link the synsets are represented by pointers: lexical and semantic. Lexical relations hold between semantically related word forms; semantic relations hold between word meanings. These relations include (but are not limited to) Hypernymy /hyponymy (superordinate/subordinate), antonymy, entailment, and meronymy /holonymy.

The purpose of this lexical resource is to create an intuitive combination of dictionary and thesaurus and to support automatic text analysis and artificial intelligence applications. Another main requirement was to develop a system that would be consistent with the knowledge acquired over the years about how human beings process language.

WordNet was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Development began in 1985. Over the years, the project received about $3 million of funding, mainly from government agencies interested in machine translation.

At the end of 2006, the database contained about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs. [6]

3.1 WordNet Structure

WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. Every synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning, such as "car pool"); different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining glosses (Definitions and/or example sentences). A typical example synset with gloss is: good, right, ripe -- (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes"). [7]

Nouns and verbs are organized into hierarchies based on the hypernymy/hyponymy relation between synsets. Additional pointers are be used to indicate other relations.

Adjectives are arranged in clusters containing head synsets and satellite synsets. Each cluster is organized around antonymous pairs (and occasionally antonymous triplets). The antonymous pairs (or triplets) are indicated in the head synsets of a cluster. Most head synsets have one or more satellite synsets, each of which represents a concept that is similar in meaning to the concept represented by the head synset. One way to think of the adjective cluster organization is to visualize a wheel, with a head synset as the hub and satellite synsets as the spokes. Two or more wheels are logically connected via antonymy, which can be thought of as an axle between the wheels.

Pertainyms are relational adjectives and do not follow the structure just described. Pertainyms do not have antonyms; the synset for a pertainym most often contains only one word or collocation and a lexical pointer to the noun that the adjective is "pertaining to". Participial adjectives have lexical pointers to the verbs that they are derived from.

Adverbs are often derived from adjectives, and sometimes have antonyms; therefore the synset for an adverb usually contains a lexical pointer to the adjective from which it is derived.

Synsets are connected to other synsets via a number of semantic relations.

These relations vary based on the type of word, and include:

- Nouns
    - *hypernyms: Y is a hypernym of X if every X is a (kind of) Y*
    - *hyponyms: Y is a hyponym of X if every Y is a (kind of) X*
    - *coordinate terms: Y is a coordinate term of X if X and Y share a hypernym*
    - *holonym: Y is a holonym of X if X is a part of Y*
    - *meronym: Y is a meronym of X if Y is a part of X*
- Verbs
    - *hypernym: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (travel to movement)*

  - *troponym: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (lisp to talk)*
  - *entailment: the verb Y is entailed by X if by doing X you must be doing Y (sleeping by snoring)*
  - *coordinate terms: those verbs sharing a common hypernym*
- Adjectives
  - *related nouns*
  - *participle of verb*
- Adverbs
  - *root adjectives*

While semantic relations apply to all members of a synset because they share a meaning but are all mutually synonyms, words can also be connected to other words through lexical relations, including antonyms (opposites of each other) and derivationally related, as well.

WordNet also provides the polysemy count of a word: the number of synsets that contain the word. If a word participates in several synsets (i.e. has several senses), then typically some senses are much more common than others. WordNet quantifies this by the frequency score: in which several sample texts have all words semantically tagged with the corresponding synset, and then a count provided indicating how often a word appears in a specific sense.

The morphology functions of the software distributed with the database try to deduce the lemma or root form of a word from the user's input; only the root form is stored in the database unless it has irregular inflected forms.

Both nouns and verbs are organized into hierarchies, defined by hypernym or IS A relationships. For instance, the first sense of the word dog would have the following hypernym hierarchy; the words at the same level are synonyms of each other: some sense of dog is synonymous with some other senses of domestic dog and Canis familiaris, and so on. Each set of synonyms (synset), has a unique index and shares its properties, such as a gloss (or dictionary) definition.

```
dog, domestic dog, Canis familiaris

  => canine, canid

    => carnivore

      => placental, placental mammal, eutherian, eutherian mammal

        => mammal

          => vertebrate, craniate

            => chordate

              => animal, animate being, beast, brute, creature, fauna

                => ...
```

At the top level, these hierarchies are organized into base types, 25 primitive groups for nouns, and 15 for verbs. These groups form lexicographic files at a maintenance level. These primitive groups are connected to an abstract root node that has been assumed by various applications that use WordNet.

In the case of adjectives, the organization is different. Two opposite 'head' senses work as binary poles, while 'satellite' synonyms connect to each of the heads via synonymy relations. Thus, the hierarchies, and the concept involved with lexicographic files, do not apply here the same way they do for nouns and verbs.

The network of nouns is far deeper than that of the other parts of speech. Verbs have a far bushier structure, and adjectives are organized into many distinct clusters. Adverbs are defined in terms of the adjectives they are derived from, and thus inherit their structure from that of the adjectives.

3.2 WordNet and Ontologies

Hypernym/hyponym relationships between noun synsets can be viewed as relations between conceptual categories. In this sense WordNet can be interpreted and used as a lexical ontology.

A well known example is to use WordNet as an ontology for determining word similarity determination for witch several algorithms have been proposed. These algorithms include the calculus of the distance between the conceptual categories of words, or the hierarchical structure of the WordNet ontology.

3.3 WordNet Limitations

WordNet can be used only as a general resource for NLP. Unlike other dictionaries WordNet does not include information about etymology, pronunciation and the forms of irregular verbs and contains only limited information about usage.

The lexicographical and semantic information is maintained in *lexicographer files*. The files are then processed by a tool called *grind* to produce the database. Because it groups similar words together under a single, general definition, the definitions WordNet provides for several individual words are not accurate.

WordNet contains a sufficient wide range of general information while it does not cover special domain vocabulary. It is designed as an underlying database for different applications. In this sense one of the main features of WordNet vocabulary is its generality.

3.4 WordNet Based Applications

WordNet has been recognized as a valuable resource in language and knowledge processing communities.

The origin of its success is to be found in its accessibility, quality and great potential in the Natural Language Processing area.

Many AI researchers view WordNet as a lexical knowledge base and use it subsequently. Knowledge processing has gained new dimensions since the development of WordNet. Its applicability has been cited in hundreds of papers and systems have been implemented based on it. One of the most comprehensive paper on WordNet applications was written in 2004 by Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro. [8]

Many groups of researchers and commercial enterprises expressed their interest in WordNet applications.

The most important fields of application are:

- Natural Language Processing
    - Information Retrieval
    - Information Extraction
    - Word Sense Disambiguation
    - Text Inference
- Machine Learning
- Knowledge Acquisition
- Multilingual WordNets

Natural Language Processing

During the last years WordNet has served as a support for the development of tools to enhance the efficiency of offline and online document searches.

Information Retrieval and Extraction

These operations are closely related to organization and representation of knowledge in natural language documents. One of the research directions is the application of artificial intelligence to information retrieval. The goal is to design automatic information retrieval systems simulating the inferential process of human reasoning.

The first times 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. 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. [8]

Basili, Roberto, Paola Velardi and Maria Teresa Pazienza[9] proposed the use of WordNet for verb classification tasks in ``Integrating general-purpose and corpus-based verb classification." In: Computational Linguistics 22 (4), 1996, pp. 559 - 568.

Also in the proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998, Basili, Roberto, Alessandro Cucchiarelli, Carlo Consoli, Maria Teresa Pazienza and Paola Velardi[10] presented ``Automatic adaptation of WordNet to sublanguages and computational tasks."

Mandala[11] 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.

Moldovan[12] preferred to use WordNet in the development of a natural language interface to optimize the precision of Internet search engines by expanding queries.

Another development direction is the multilingual area. Regarding this topic Basili, R., R. Catizone, L. Padro, M.T. Pazienza, G. Rigau, A. Setzer, N. Webb and F. Zanzotto[13] presented “Knowledge-Based Multilingual Document Analysis” In the *Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks"*, Taipei, August 2002.

Word Sense Disambiguation

Disambiguation is without any doubt the most explored WordNet application. In this field WordNet has been used to enhance the information retrieval efficiency and several interesting research areas had emerged:

- The Development of classifiers able to combine a neuronal network to process subject context and a network to process local context.
- The exploitation of Bayesian networks able to establish lexical relations with WordNet as a source of knowledge.
- Support for the development of a computational similarity model to add on-line semantic representation to the statistical corpus.

During the years WordNet has proved its worth as an really good and robust methodological element to disambiguate the meaning of words in information extraction systems.

Projects have been launched to disambiguate nouns in English 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 optimizing information retrieval.

One example is the IWA/H project for building an ontological framework. The framework should be able to disambiguate search criteria via mixed knowledge representation technique systems (ARPA KRSL).

Other examples include tools such as Oingo and SimpliFind, two Internet products that avoid ambiguity in natural language searches by using the WordNet lexicon 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[14].

An important topic is the contextual word sense extraction. This topic was presented by Basili, Roberto, M. DellaRocca and Maria Teresa Pazienza[15] in “Contextual word sense tuning and disambiguation.” In: *Applied Artificial Intelligence* 11 (3), 1997, pp. 235 - 262.

Text inference

Text inference refers to extracting relevant, new information from a text.
We have a great ability to perform correct inference from text or speech. In order to achieve this we use a great deal of world knowledge and we can focus our thoughts and filter out irrelevant facts.

Machine Learning

Machine Learning based Natural Language Processing is a very active and multidisciplinary area of research.

Machine Learning is about automatically acquiring knowledge from a concrete data domain using computers. The main goal is to obtain a description of the concept in a representation language that explains observations and helps predicting new instances of the same distribution.

Machine Learning is applied to solve tasks as:

- Reacting to environmental inputs
- Learning concepts from data
- Solving lexical and structural ambiguity problems

Knowledge Acquisition

Knowledge acquisition has become a major area of artificial intelligence and cognitive science research. Although knowledge acquisition and machine learning have been considered as separate subfields of AI, there is a tendency for the two fields to come together.

Typical methods of knowledge acquisition tend to need an expert so that systems can be explained and problems be solved. These experts can sometimes be unattainable or unhelpful. Expert systems are then needed to find and solve problems without the need for knowledge acquisition techniques to be used on the experts. Automating this knowledge acquisition phase then becomes important. When there are large amounts of data the expert system can study this knowledge and generate rules that define the knowledge. The knowledge can then be refined and the system used to give solutions.

One of the most important tasks of Knowledge Acquisition (KA) is Ontology Building. Ontologies can be viewed as taxonomic catalogues of concept types and relation types.

Most of the natural language concepts and relations may appear in documents which are sources of expertise (technical documents, interview transcriptions, etc.) and a great number of these concepts might have to be clusterized or classified for the KA process. One of the possible approaches is to use Ontologies build on the WordNet lexical resource. [7]

This approach is useful both in Document Clustering and Information Extraction.

Multilingual WordNets

One of the most relevant activities has been the development of EuroWordNet[16], a project based on WordNet structure whose ultimate purpose is to develop multilingual databases with wordnets for several European languages *(Dutch, Italian, Spanish, German, French, Czech and Estonian)*.

Each wordnet adopts an autonomous lexicalization structure and all are interconnected through an interlinguistic index, for which relations have been added and modified and new levels identified in WordNet.

The wordnets are structured in the same way as the American wordnet for English (Princeton WordNet, Miller et al 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The database can be used, among others, for monolingual and cross-lingual information retrieval, which was demonstrated by the users in the project.

The EuroWordNet project was completed in the summer of 1999. The design of the database, the defined relations, the top-ontology and the Inter-Lingual-Index are now frozen. Nevertheless, many other institutes and research groups are developing similar wordnets in other languages (European and non-European) using the EuroWordNet specification. If compatible, these wordnets can be added to the above database and, via the index, connected to any other wordnet. The EuroWordNet format is defined by the EuroWordNet Database Editor Polaris. A specification can be found in the user-manual of the database. To our knowledge, wordnets are currently developed for Swedish, Norway, Danish, Greek, Portuguese, Basque, Catalan, Romanian, Lithuanian, Russian, Bulgarian, Slovenian (BalkanNet).

The design of EuroWordNet makes it possible to precisely describe the lexicalization of a language given a conceptual space. What this implies is best illustrated with an example. Consider, for example, all the words that are related to body parts. All the wordnets will share the top-ontology concepts **Part** and **Living** but each language has different lexicalizations for body parts due to the differences between languages.

3.5 Trends

Trends are difficult to determine and evaluate in view of the application oriented dimension that underlies WordNet's success.

In the recent publications we can distinguish a number of trends in WordNet use. [8]:

1. Development of interlinguistic indices for concept equivalence. Also a research area is the integrated access to information based on multiple database access systems.
2. Tool for the optimization of the retrieval capacity of existing systems: search engines natural language interfaces; automatic generation semantic disambiguation tools, creation of knowledge summaries from expanded queries.
3. Support for information classifiers based on grammatical categorizations.
4. Audio-visual and multi-media information retrieval systems.
5. Ontology construction for the Semantic Web.

References

Knowledge Representation

1. Ludwig Wittgenstein, Das logische Bild der Tatsachen ist der Gedanke, Tractatus Logico-Philosophicus, 1921
2. Trevor JM Bench-Capon, Knowledge Representation: An Approach to Artificial Intelligence - Academic Press, London, 1990

Semantic networks

3. Ross Quillian, Semantic memory PhD Thesis, 1966
4. WordNet, http://wordnet.princeton.edu
5. Birger Hjørland, Semantic Network, 2006

WordNet

6. WordNet 3.0 official statistics - http://wordnet.princeton.edu/man/wnstats.7WN
7. George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, Introduction to WordNet: An On-line Lexical Database, 1993
8. Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro, WordNet Applications, 2004
9. Basili, Roberto, Paola Velardi and Maria Teresa Pazienza. ``Integrating general-purpose and corpus-based verb classification." In: Computational Linguistics 22 (4), 1996, pp. 559 - 568.
10. Basili, Roberto and Alessandro Cucchiarelli and Carlo Consoli and Maria Teresa Pazienza and Paola Velardi. ``Automatic adaptation of WordNet to sublanguages and computational tasks." In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998.
11. Mandala, Rila, Tokunaga, T., Tanaka, Hozumi, O., Akitoshi, Satoh, K.: Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri, 1998
12. Moldovan, D. I. and Mihalcea, R.: Using WordNet and lexical operators to improve Internet searchers, IEEE Internet Computing, 2000
13. Basili, R. and R. Catizone and L. Padro and M.T. Pazienza and G. Rigau and A. Setzer and N. Webb and F. Zanzotto ``Knowledge-Based Multilingual Document Analysis" In: Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks", Taipei, August 2002. http://www.cs.ust.hk/~hltc/semanet02/pdf/basili.pdf
14. Eichmann, David: Balancing the Need for Knowledge and Nimbleness in Transportable Agents, 1996
15. Basili, Roberto, M. DellaRocca and Maria Teresa Pazienza. ``Contextual word sense tuning and disambiguation." In: Applied Artificial Intelligence 11 (3), 1997, pp. 235 - 262.
16. EuroWordNet - Building a multilingual database with wordnets for several European languages, 1999 - http://www.illc.uva.nl/EuroWordNet/