Ontology Learning for Portuguese

Student: Hugo Gonçalo Oliveira* (hroliv@dei.uc.pt)
Supervisor: Paulo Gomes (pgomes@dei.uc.pt)
Doctoral Program in Information Science and Technology
Beginning date: October 2008
Foreseen conclusion: January 2012

Cognitive and Media Systems Groups
Centre for Informatics and Systems
University of Coimbra

Abstract. Having in mind the importance that lexical resources play nowadays in natural language processing (NLP), this research aims the automatic creation of a lexical ontology for Portuguese. For its purpose, patterns in textual resources will be exploited in order to acquire relations between concepts. The ontology will be evaluated and then made available for the community.

Key words: Natural Language Processing, Natural Language Applications, Information Extraction, Ontologies

1 Introduction

There is a growing number of applications that perform tasks where lexico-semantic resources are needed. Tasks that go from automatic generation of text to intelligent search and machine translation, as well as writing aids demonstrate that NLP is becoming more and more dependent on semantic information.

Lexical ontologies are models aiming to represent the lexical structure and thus, the meaning of a language, as opposing to terminologies or domain ontologies, whose purpose is to describe specific topics or domains. While for English WordNet [1] was established as the standard model of a lexical database, the picture is quite different for other languages, like Portuguese, where similar resources [2][3] are currently in development and not publicly available for download. In order to avoid time-consuming human work in its construction and maintenance, our goal is to automate the creation of a lexical ontology for Portuguese, that in the future will be in the public domain.

We start by stating the main goals of this research (Section 2) and introducing some existing resources, similar to the one we want to develop (Section 3). Our work plan is presented and discussed along with related work (Section 4) on the extraction of relations from machine readable dictionaries (MRDs) and corpora, and on the evaluation of ontologies. Before concluding (Section 6) work that we have been doing is also referred (Section 5).

* Supported by the FCT grant SFRH/BD/44955/2008

2 Research Goals

The main goal of this research is to design and develop the tools needed to create a lexical ontology for Portuguese, by semi-automatic means. We intend the resulting ontology to be in the public domain and freely available for download, so that in a near future it can be used by the Portuguese NLP community and also by other researchers that need Portuguese lexical knowledge in their work.

While there is intensive labour involved in manually encoding lexical entries, lexical capabilities of NLP systems will always be weak [4]. Handcrafting ontologies is impractical and undesirable and we should take advantage of available NLP tools in order to automate part of this task, reducing the need of manual input [5].

In a lexical database, concepts are organised in a network and relate with other concepts by means of semantic relations. These relations are present in text and can be identified by the usage of specific patterns. In Table 1 common relations and examples of textual patterns (for Portuguese), associated with them, are shown. Different textual sources will be exploited in order to populate the ontology: first MRDs [6] for acquiring general knowledge and then corpora to complete specific lexical gaps.

Relation	Example pattern
Hypernymy	`tipo|género|classe|forma de`
Meronymy	`parte|membro de`
Causation	`causado|provocado|originado por`
Purpose	`usado|utilizado para`

Table 1. Examples of patterns indicating semantic relations.

We also intend to evaluate the ontology (or the ontologies), preferably by scalable semi-automatic means, but we do not discard the manual evaluation of a subset of the results.

One of the challenges involved is that for Portuguese, as for the majority of other non-English languages, the amount of existing NLP resources has no possible comparision with the amount of NLP resources for English so we will have to come up with new ideas or recycle old ones.

3 Similar Resources

Princeton WordNet [1] is probably the most important reference when it comes to lexical databases in English. It is freely available and widely used in NLP research. In the WordNet's lexicon, the words are clearly divided into nouns, verbs, adjectives, adverbs and functional words. The basic structure in WordNet is the *synset*, which is a set of synonym words that can be used to represent one concept. The *synsets* are organised in a network of semantic relations, such

as hyponymy and meronymy (between nouns) and troponymy and entailment (between verbs).

There are attempts for creating (from scratch) a "wordnet" for portuguese, namely WordNet.PT [3] and WordNet.BR [2], but they are both still in a development stage. Besides these, Portuguese is one of the languages aligned with Princeton WordNet in the scope of the MultiWordNet project [7]. There is also Tep [8], which is an electronic thesaurus for Brazilian Portuguese, developed under the principles of WordNet and freely available for download.

MindNet [9] is a lexical knowledge base created automatically not only from MRDs but also from encyclopedias, and free text, with the help of a broad-coverage parser. MindNet contains a long set of relations, including Hypernymy, Causation, Meronymy, Manner, Location and many more. One interesting functionality offered by MindNet is the identification of "relation paths" between words. Each path is automatically weighted according to its salience and can be useful to determine the similarity between two words.

Another kind of lexical resource is FrameNet [10], which constitutes a network of relations between semantic frames, manually extracted from a systematic analysis of semantic patterns in corpora. Each frame corresponds to a concept and describes an object, a state or an event by means of syntactic and semantic relations of the lexical item that represents that concept.

4 Work Plan

In this section, the most important phases of our research are presented along with related work.

4.1 Extraction of Relations from MRDs

The process of using MRDs in NLP started forty years ago with early works of Calzolari [11], for Italian and Amsler [12] for English. MRDs were analysed and, taking advantage of the simple structure of the definitions and of the restricted vocabulary used, procedures were developed to extract and structure lexical information. Dictionary definitions often have two distinct parts: a *genus*, that identifies the superordinate concept and a *differentia*, where the properties for the distinction between the instance of the superordinate concept and other instances of the same concept can be found.

Procedures were developed to extract semantic hierarchies from a MRD [13] based on string patterns and specific grammars were proposed for parsing definitions of particular dictionaries and produce semantic structures [14]. Other authors [15] used broad-coverage parsers to extract semantic information from dictionary text, claiming they were better suit to capture the features in the *differentia*, even though this was not consensual in the community [16]. More recently, Nichols et al. [17] and O'Hara [18] also used MRDs for the automatic extraction of lexical ontologies.

In any case, one of the main reasons for using dictionaries and not (only) running text is because MRDs are highly structured, they are a substancial source of general lexical knowledge [4], and the "authorities" of word sense [19]. Dictionaries have thus been exploited for several purposes, such as parsing or word sense disambiguation (WSD), but to our knowledge they have not been converted into an independent resource of its own before MindNet [20]. MindNet can therefore be claimed to be a kind of independent (dictionary-based) lexical ontology in a way that previous work was not.

PAPEL [21] is a lexical resource for Portuguese, consisting of relations between terms, extracted after processing the definitions of a major general dictionary. It contains about 200,000 relations organised into main groups, that can be divided into sub-relations, according to the grammatical category of the arguments (see Table 2).

Group	**Name**	**Args.**	**Qnt.**	**Examples**
Synonymy	SINONIMO_DE	*same*	80,432	(*flexível*, *moldável*)
Hypernymy	HIPERONIMO_DE	sub,sub	63,455	(*planta*, *salva*)
Meronymy	PARTE_DE	sub,sub	14,453	(*cauda*, *cometa*)
	PARTE_DE_ALGO_COM_PROP.	sub,adj	3,715	(*tampa*, *coberto*)
	PROPRIEDADE_DE_ALGO_PARTE_DE	adj,sub	962	(*celular*, *célula*)
Cause	CAUSADOR_DE	sub,sub	1,125	(*fricção*, *assadura*)
	CAUSADOR_DE_ALGO_COM_PROP.	sub,adj	16	(*paixão*, *passional*)
	PROPRIEDADE_DE_ALGO_CAUSADOR_DE	adj, sub	5,15	(*reactivo*, *reacção*)
	ACCAO_QUE_CAUSA	v,sub	6,424	(*limpar*, *purgação*)
	CAUSADOR_DA_ACCAO	sub,v	39	(*gases*, *fumigar*)
Purpose	FINALIDADE_DE	sub,sub	2,095	(*defesa*, *armadura*)
	FINALIDADE_DE_ALGO_COM_PROP.	sub,adj	23	(*reprodução*, *reprodutor*)
	ACCAO_FINALIDADE_DE	v,sub	5,640	(*fazer_rir*, *comédia*)
	ACCAO_FINALIDADE_DE_ALGO_COM_PROP.	v,adj	255	(*corrigir*, *correccional*)
	MANEIRA_POR_MEIO_DE	adv,sub	1,433	(*timidamente*, *timidez*)
Place	LOCAL_ORIGEM_DE	sub,sub	768	(*Japão*, *japonês*)

Table 2. Some relations of PAPEL.

In PAPEL, grammars were manually created for the extraction of the predefined relations, having in mind the specific structure of the definitions in the dictionary. To extract the relations, a chart parser processes the definitions according to the grammars and, if the definition suits the rules, a derivation tree is generated. Finally, for each grammar, the extraction tool selects the better tree and outputs eventual relations (identified by the labels of the tree nodes) between words in the definition and the defined word.

It is our intention to use PAPEL as the base for our ontology. Its structure and relations will be analysed in order to improve the grammars, the extraction tools and thus the quality of the relations. We are also planning, as suggested by Ide and Véronis [22], to adapt the extraction procedure to other MRDs in order to refine our results. One possible MRD to exploit is the Portuguese version of Wiktionary[1], a collaborative dictionary maintained by the Wikimedia Foundation.

[1] http://pt.wiktionary.org/

4.2 Resource Structure

Decisions about the resource structure will be made after having the relation set extracted from MRDs. One idea is to adopt a "wordnet-like" structure, where synonym words are included in the same synset, and the relations occur between synsets. To achieve this kind of structure, WSD techniques would be needed to identify possible different senses of a word. We do not expect this to be easy, since there is no consensus concerning WSD [19] and it is very dependent on the purpose [23]. We have however some ideas on how to get useful hints to accomplish WSD: the sense division in dictionaries; sentences where the words occur, preferably sentences where we know what sense is being used (e.g. example sentences in Tep [8]); or exploitation of the (ambiguous) network structure.

Another possibility is to create a structure similar to Mindnet, where the sense division in the MRDs is used to define the various senses a word can have. The main structure would be the word, that would contain its possible grammatical categories and senses. Relations would occur between a word sense and a word structure. This approach would require fairly less WSD and would eventually have to deal with word sense ambiguation [24], in order to group related word senses.

4.3 Extraction of Relations from Corpora

As it is argued by several authors [16], in order to find terms and expressions that are not defined in MRDs, we must turn to other textual resources, like textual corpora, that should be viewed as the main source of domain-specific information [5]. So, through the lines of Hearst [25], our idea is to develop tools to extract relations from corpora and use them either to enrich the main ontology in specific domains or to create new domain ontologies based on the texts.

Work on the discovery of relations from text using large corpora became the paradigm in ontology construction after Hearst's [16] seminal work, where an automatic method to discover lexico-syntatic patterns, used for the acquisition of hyponyms, is proposed. Many works were inspired by Hearst's work, not only for the hyponymy relation [26], but also for other relations like meronymy [27] or causality [28].

Still concerning automatic extraction of relations from text, Banko et al. [29] propose a new paradigm where the system makes a single data-driven pass over a corpus and extracts a large set of relational tuples, without requiring any human input. The output triples were later used to create an ontology [30], with the help of WordNet that was used as a map of concepts.

We will try to adapt the tools used to extract relations from MRDs to corpora but we are aware that extraction from unrestricted text will be more difficult because this kind of text is not structured, its vocabulary is not controlled and it may contain several features like metaphors and anaphora. Experiences using annotated corpora (e.g. CETEMPúblico [31]) will be made and we will also devise using pos-taggers or syntactic annotators in the extraction procedure.

4.4 Evaluation of Ontologies

For domain ontologies, Brank et al. [32] divide evaluation approaches into four groups: (i) performed by human subjects; (ii) comparison with a golden standard; as for coverage, (iii) comparison with a collection of documents about a domain covered by the ontology; (iv) complete some task that uses the ontology.

Although the most reliable in the end, human evaluation does not take advantage of computer programs and relies heavily on time consuming work from domain specialists. We will try to avoid it, but we do believe that we might need it sometime during our research to have a clearer notion of the quality of our results. In several works [33] [17] small scale human evaluations of relations were performed and common statistical techniques were used to estimate the representativeness of the evaluated set. As for having a specialist, we do believe it to be dispensable: if the resource is made to suit the community's needs, it could as well be evaluated by the community. So, one idea would be to have potential users classifying the extracted relations in an online inquire or interactive game.

Still, we will try to automate the evaluation procedure as much as possible, also because it will enable an easier repetition of the evaluation procedure. The ontology can be compared with some other resource (e.g. another ontology) that is known to be correct, eventually because it was manually created by specialists. For example, WordNet was used as a golden resource in several works [16][17] that, besides evaluating the relations, were able to find gaps in Wordnet.

But if this may be OK to validate a particular automatic method, it is obviously of little practical interest, because one expects to be creating new ontologies, not recreating existing ones. So, while the approach of compiling a human resource is commonly followed in joint evaluations, for example ReRelEM [34], where system's capabilities to recognise semantic relations between named entities were evaluated, it can only encompass a few examples. Besides, the only possible golden resource we are aware that we can be freely downloaded and used in our evaluation is Tep. It is for sure a good option for evaluating synonymy, but it does not have the other relations.

The two other approaches are hardly adapted to lexical ontologies. Yet they are interesting possibilities to evaluate how well the knowledge on some domain can be enriched, after processing domain-specific corpora. The third method consists of finding how adequate a particular ontology is for representing the knowledge contained in a collection of documents, as in Brewster et al.'s [35] measurement of the fit between an ontology and a corpus: after identifying salient terms in a domain corpus and looking for them in a same domain ontology, the fit is proportional to the number of terms found in both corpus and ontology. The problem is that we cannot define a clear set of salient terms for general language, so this method cannot be applied to a lexical ontology. In the last approach, external or task-based, indirect evaluation is performed by assessing the performance of an application which uses the ontology to do some task. Porzel and Malaka [36] proposed this approach aiming at evaluating ontologies with respect to the fit of the vocabulary, the fit of the taxonomy and the adequacy of non-taxonomic semantic relations.

The truth is that evaluation is not usual when it comes to dictionaries and lexical ontologies, because these resources are typically the result of manual work by experts and thus are not prone to errors. Nevertheless, we are aware of independent evaluations, for instance an automatic evaluation of the Wordnet synsets, with the help of a dictionary they used to obtain synonyms and hyponyms [37]. For Mindnet [9], created automatically, an (incomplete) evaluation of the quality of the semantic relations is referred, but the description of the evaluation process does not go very far. One comment made is that the quality varies according to the relation type.

Considering all the approaches, it is also possible to combine some of them, like the idea followed in the first evaluation of PAPEL, briefly described in Section 5.

4.5 Deployment

After reaching an adequate level of quality we intend to make the resource and the tools publicly available, together with user documentation.

To extend the potential utilisation scenarios we are devising to export the resource to several data representation formats. For example, there are many Semantic Web [38] applications based on RDF/OWL [39] [40] models, because these are the W3C standard description languages for the Semantic Web. Additionally, these languages ease the browsing and visualisation of ontologies and have other useful features like the possibility of creating rules for inference of new relations and reasoning, so there is a strong possibility of developing a RDF/OWL representation of our ontology.

5 Current work

Since the beginning of our work we have been analysing the results of PAPEL in order to find out problems that can be corrected, for example in the grammars, to improve the quality of the relations.

In order to validate the relations semi-automatically, we have also developed a testing system where synonymy relations are evaluated using the thesaurus Tep as a golden resource. As for other relations, they are rendered into natural language and the obtained patterns are searched in textual corpora (more precisely CETEMPúblico [31]), in a similar fashion to what Etzioni et al. [41] have done to evaluate their hyponymy relations using the Web. The number of patterns found in the corpus gives us an idea of the quality of the relations.

Shifting to the Semantic Web, we have converted PAPEL into an OWL model, and then developed an interface to help us visualising and browsing through OWL networks, VisuOWL[2]. This tool (see Figure 1) has revealed to be very useful for getting a clearer idea about the results of PAPEL and also for debugging.

[2] Available for download through http://code.google.com/p/visuowl/

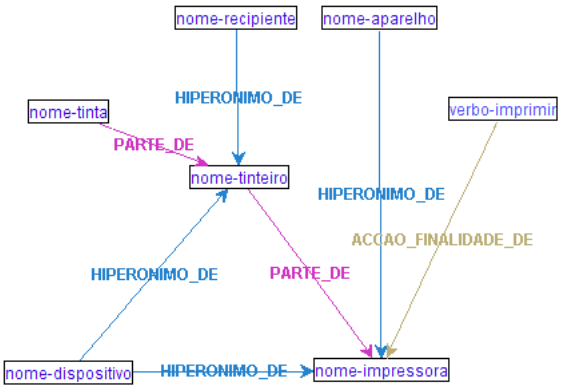

Fig. 1. The VisuOWL tool.

6 Concluding remarks

This research intends to create a lexical ontology for Portuguese by semi-automatic means. The goals of the research were presented, and so has the plan we intend to follow to obtain, organise and evaluate the results. In the end of the research the resulting resource should be made public to all the NLP community, as well as the tools developed, and we hope that in a near future it might be used by researchers and developers that work with Portuguese. There is still a long way to go, but we believe results will come and future Portuguese NLP applications will have a useful resource to complement them and increase their potential.

References

1. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998)
2. Dias da Silva, B.C., Oliveira, M., Moraes, H.: Groundwork for the Development of the Brazilian Portuguese Wordnet. In Mamede, N., Ranchhod, E., eds.: Proc. Advances in Natural Language Processing: 3rd Intl. Conference. LNAI, Berlin/Heidelberg, Springer Verlag (2002) 189–196
3. Marrafa, P.: Portuguese wordnet: general architecture and internal semantic relations. DELTA **18** (2002) 131–146
4. Briscoe, T.: Lexical issues in natural language processing. In Klein, E., Veltman, F., eds.: Natural Language and Speech: Symposium Proc. Springer, Berlin, Heidelberg (1991) 39–68
5. Brewster, C., Wilks, Y.: Ontologies, taxonomies, thesauri: Learning from texts. In Deegan, M., ed.: Proc. Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop, London, UK, Centre for Computing in the Humanities, Kings College (5-6 February 2004)
6. Wilks, Y., Fass, D., ming Guo, C., Mcdonald, J.E., Plate, T., Slator, B.M.: Machine tractable dictionaries as tools and resources for natural language processing.

In: Proc. 12th Conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1988) 750–755
7. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: 1st Intl. Conference on Global WordNet. (2002)
8. Maziero, E., Pardo, T., Di Felippo, A., Dias-da Silva, B.: A base de dados lexical e a interface web do tep 2.0 - thesaurus eletrônico para o português do brasil. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL). (2008) 390–392
9. Vanderwende, L., Kacmarcik, G., Suzuki, H., Menezes, A.: Mindnet: An automatically-created lexical resource. In: HLT/EMNLP, The Association for Computational Linguistics (2005)
10. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proc. 17th Intl. conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1998) 86–90
11. Calzolari, N., Pecchia, L., Zampolli, A.: Working on the italian machine dictionary: a semantic approach. In: Proc. 5th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1973) 49–52
12. Amsler, R.A.: A taxonomy for english nouns and verbs. In: Proc. 19th annual meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1981) 133–138
13. Chodorow, M.S., Byrd, R.J., Heidorn, G.E.: Extracting semantic hierarchies from a large on-line dictionary. In: Proc. 23rd annual meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1985) 299–304
14. Alshawi, H.: Analysing the dictionary definitions. Computational lexicography for natural language processing (1989) 153–169
15. Montemagni, S., Vanderwende, L.: Structural patterns vs. string patterns for extracting semantic information from dictionaries. In: Proc. 14th Conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1992) 546–552
16. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. 14th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1992) 539–545
17. Nichols, E., Bond, F., Flickinger, D.: Robust ontology acquisition from machine-readable dictionaries. In Kaelbling, L.P., Saffiotti, A., eds.: IJCAI, Professional Book Center (2005) 1111–1116
18. O'Hara, T.P.: Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions. PhD thesis, NMSU CS (August 2005)
19. Kilgarriff, A.: "I don't believe in word senses". Computing and the Humanities **31**(2) (1997) 91–113
20. Richardson, S.D., Dolan, W.B., Vanderwende, L.: Mindnet: Acquiring and structuring semantic information from text. In: COLING-ACL. (1998) 1098–1102
21. Gonçalo Oliveira, H., Gomes, P., Santos, D., Seco, N.: PAPEL: a dictionary-based lexical ontology for Portuguese. In Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P., eds.: Computational Processing of the Portuguese Language, 8th Intl. Conference, Proceedings (PROPOR 2008). Volume 5190., Springer Verlag (2008) 31–40
22. Ide, N., Véronis, J.: Refining taxonomies extracted from machine readable dictionaries. In Hockey, S., Ide, N., eds.: Research in Humanities Computing 2. (1993)
23. Wilks, Y.: Is word sense disambiguation just one more nlp task? Computers and the Humanities **34** (2000) 235–243
24. Dolan, W.B.: Word sense ambiguation: clustering related senses. In: Proc. 15th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1994) 712–716

25. Hearst, M.: Automated discovery of wordnet relations. In Fellbaum, C., ed.: WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press (1998) 131–153
26. de Freitas, M.C.: Elaboração automática de ontologias de domínio: discussão e resultados. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro (Janeiro 2007)
27. Berland, M., Charniak, E.: Finding parts in very large corpora. In: Proc. 37th Annual Meeting of the ACL on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1999) 57–64
28. Girju, R., Moldovan, D.: Text mining for causal relations. In Haller, S.M., Simmons, G., eds.: Proc. 15th Intl. Florida Artificial Intelligence Research Society Conference (FLAIRS). (2002) 360–364
29. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In Veloso, M.M., ed.: Proc. Intl. Joint Conference on Artificial Intelligence (IJCAI). (2007) 2670–2676
30. Soderland, S., Mandhani, B.: Moving from textual relations to ontologized relations. In: Proc. AAAI Spring Symposium on Machine Reading. (2007)
31. Rocha, P.A., Santos, D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In das Graças Volpe Nunes, M., ed.: V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR), São Paulo, ICMC/USP (2000) 131–140
32. Brank, J., Grobelnik, M., Mladenic', D.: A survey of ontology evaluation techniques. In: Proc. Conference on Data Mining and Data Warehouses (SiKDD). (2005)
33. Richardson, S., Vanderwende, L., Dolan, W.: Combining dictionary-based and example-based methods for natural language analysis. In: Proc. 5th Intl. Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan (1993) 69–79
34. Freitas, C., Santos, D., Mota, C., Gonçalo Oliveira, H., Carvalho, P.: Detection of relations between named entities: report of a shared task. In: Semantic Evaluations: Recent Achievements and Future Directions (SEW), NAACL-HLT Workshop. (2009)
35. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data-driven ontology evaluation. In: Proc. Language Resources and Evaluation Conference (LREC), Lisbon, Portugal, European Language Resources Association (2004) 164–168
36. Porzel, R., Malaka, R.: A task-based approach for ontology evaluation. In Buitelaar, P., Handschuh, S., Magnini, B., eds.: Proc. ECAI 2004 Workshop on Ontology Learning and Population, Valencia, Spain (2004)
37. Raman, J., Bhattacharyya, P.: Towards automatic evaluation of wordnet synsets. In Tancs, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P., eds.: Proceedings of the 4th Global WordNet Conference (GWC 2008), Szeged, Hungary, University of Szeged, Department of Informatics (2008)
38. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (May 2001)
39. Miller, E., Manola, F.: RDF primer (2004) Published: W3C Recommendation.
40. McGuinness, D.L., van Harmelen, F.: OWL web ontology language overview (2004) Published: W3C Recommendation.
41. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: an experimental study. Artificial Intelligence **165**(1) (2005) 91–134