

Enriching the Romanian WordNet Using Semi-automatically Identified Hyponymic Patterns

Verginica Barbu Mititelu

Romanian Academy Research Institute for Artificial Intelligence

13, Calea 13 Septembrie, Bucharest 050711, Romania

vergi@racai.ro

Dan Ștefănescu

Romanian Academy Research Institute for Artificial Intelligence

13, Calea 13 Septembrie, Bucharest 050711, Romania

danstef@racai.ro

Alexandru Ceașu

Romanian Academy Research Institute for Artificial Intelligence

13, Calea 13 Septembrie, Bucharest 050711, Romania

aceausu@racai.ro

Abstract

This paper presents the algorithm used for the semi-automatic identification of Romanian hyponymic patterns, the precision of these patterns, which we further use for identifying on the web and filtering instances or Named Entities to be introduced in the Romanian WordNet.

1 Introduction

The development of the Romanian WordNet (RoWN) started in the BalkaNet project (Tufiș et al., 2004). Since then it has been being enriched so that at the moment of this writing, RoWN contains 55,985 synsets (with 49,439 lemmas) classified according to the DOMAINS3.1 taxonomy (Bentivogli et al., 2004), aligned to the SUMO&MILO ontology concepts (Niles and Pease, 2001), labeled with subjectivity markups (Tufiș, 2009) of the SentiWordNet type (Esuli and Sebastiani 2006). All words in glosses have been lemmatized, tagged and parsed (with dependency links).

RoWN has been used as a lexical resource for most of the system applications developed in our Institute: TREQ-AL (Tufiș et al., 2003), COW-AL (Tufiș et al., 2006a), SynWSD (Ion and Tufiș, 2007), QA systems (Pușcașu et al., 2006, Ion et al. 2008, Ion et al., 2009). The very good results of these systems in competitions are, in great part, due to the quality of the linguistic resource they use, i.e. the RoWN.

This paper presents the way we enriched the RoWN with instances from the geographical domain using hyponymic patterns. We present the methodology for identifying these patterns for English and Romanian (section 2), alongside with their precision (section 3). We compare the

Romanian hyponymic patterns with the English ones. Section 4 describes the methodology for selecting the target concepts from the geography domain for which we identified instances on the Internet (section 5). For refining the results, we process the snippets extracted by Google (section 6). We present the results in section 7, the related work in section 8, and then conclude our paper, envisaging further work.

2 Patterns Identification

In (Barbu Mititelu 2008) we presented the methodology for semi-automatic identification of hyponymic patterns. For English we ran an experiment on over 38 million words from British National Corpus (BNC) from which we automatically extracted those sentences containing nouns in hyponymy relation, which we recognized using Princeton WordNet 2.0¹. We automatically grouped the extracted sentences according to the similarity of the lexical material (i.e. lemmas) between the hyponym and its co-occurring hypernym, thus resulting groups of examples with identical lexical material between the hyponym and its hypernym. These examples were then manually inspected to extract the hyponymic patterns.

For Romanian we did not have such a large corpus at that moment, so we proceeded to two different methods for identifying the patterns: one was the translation of the English ones found as described above, and the second was the running of the same experiment as for English but on a small Romanian corpus (1 million words) and using the RoWN (a version having more 46,000 synsets) as source of hyponym-hypernym pairs.

¹ In version 2.0 of Princeton WordNet the INSTANCE-OF relation does not exist; instances are recorded as hyponyms.

A comparison between the English and Romanian hyponymic patterns shows that different languages have the tendency to “lexicalize” this lexicon-semantic relation in similar ways (the intersection of the sets of patterns in the two languages contains 70% of the English ones and 66.6% of the Romanian ones).

3 Patterns Testing

The English patterns were tested in order to establish their precision on a file (of 7 million words) from BNC. The most representative results are in Table 1:

English Pattern	Precision (%)
NP <i>other than</i> NP	100
NP <i>especially</i> NP	100
NP <i>principally</i> NP	100
NP <i>usually</i> NP	100
NP <i>such as</i> NP	99.2
NP <i>in particular</i> NP	92.3
NP <i>e(.)g(.)</i> NP	91.4
NP <i>become</i> NP	91
NP <i>another</i> NP	87
NP <i>notably</i> NP	86.8
NP <i>particularly</i> NP	84.6
NP <i>except</i> NP	84.6
NP <i>called</i> NP	81.5
NP <i>like</i> NP	81.3
NP <i>including</i> NP	80.6
NP <i>mainly</i> NP	75
NP <i>mostly</i> NP	70.8

Table 1. Precision of English hyponymic patterns

For the present study we tested the Romanian hyponymic patterns on a sub-corpus of the OPUS corpus², namely the EMEA (European Medicines Agency) documents (11,914,802 tokens). However, it abounds in repeated expressions and a specialized vocabulary (Tiedemann, 2009). The results are in Table 2.

Romanian Pattern	Precision (%)
NP <i>chiar și</i> NP (“even”)	100
NP <i>de obicei</i> NP (“usually”)	100
NP, <i>ci (și/doar)</i> NP (“but also”)	100
NP <i>în special</i> NP (“especially”)	96.88
NP <i>precum</i> NP (“such as”)	94.83
NP <i>cum ar fi</i> NP (“such as”)	93.75

² <http://www.let.rug.nl/~tiedemann/OPUS/>

NP <i>(în) afară de</i> NP (“except”)	92.11
NP <i>și (orice) alt</i> NP (“and (any) (an)other”)	90.1
NP <i>fi un</i> NP (“be a”)	87.98
NP <i>sau alt</i> NP (“or (an)other”)	86.96
NP <i>mai ales</i> NP (“especially”)	85.71
NP <i>alt decât</i> NP (“other than”)	85.71
NP <i>sine numi</i> NP (“be called”)	84
NP <i>inclusiv</i> NP (“including”)	83.51
NP <i>de exemplu</i> NP (“for example”)	79.57
NP <i>fi considerat</i> NP (“be considered”)	79.17
NP <i>care fi</i> NP (“that be”)	74.12
NP, <i>adică</i> NP (“namely”)	66.66
NP <i>cu excepția</i> NP (“except”)	54.55
NP <i>și (tot) celălalt</i> NP (“and (all) other”)	54.29

Table 2. Precision of Romanian hyponymic patterns

Comparing the precisions of the Romanian and the English patterns, we notice that many of them are quite similar (or even identical). This is a further proof that different languages tend to lexicalize hyponymy in similar ways.

4 Selection of Target Concepts

To proceed to the semi-automatic identification of instances in the geographical domain to include in the RoWN, we needed to select from the RoWN those concepts whose daughters are instances we are interested in identifying. The RoWN was aligned to the XML version of PWN 2.1 (Tufiş et al., 2006b). We extracted from it those synsets that have synsets in INSTANCE-OF relation with them (i.e. parent synsets whose daughters are in INSTANCE-OF relation with them).

From the resulted synsets we selected only those that belong to the geography domain: 38 synsets containing 54 literals.

5 Extracting Snippets with Geographical NEs

We introduced these literals in the position normally occupied by the hypernym in five Romanian hyponymic patterns with high precision (see Table 3), thus obtaining some “seeds”. For each seed, we extracted its occurrences on the Internet using a Google Screen Scraper C# library. It allows one to input expressions into the Google Search Engine and get all the results of the search into structured objects, which can be used by different applications.

Our seed is transformed into an URL link, after which an *HttpWebRequest* is performed for that link. The request goes to Google Search Engine, which returns an html page containing results corresponding to a user input for the expression, with search parameters embedded in the link. An *HttpWebResponse* gets the html results page. An *HtmlParser* parses the page and builds an object containing all the info in the page: the html source code, the URL of the page, number of occurrences and list of items. An item corresponds to a hit of the Google Engine. It contains: the title of the hit document, the snippet (the object of our interest), the type of the document (html, pdf, etc.) and the links to the cached document and similar pages.

6 Snippets Processing

For improving the results, we decided to process the snippets extracted by Google from the Internet. Two important sources of noise in these snippets are homonymy and disregard of sentence boundary. Google does not morphologically disambiguate and segment the texts in which it searches. Thus, it returns snippets that are sometimes useless: graphical identity of different words (homography or even homonymy), words in vicinity but belonging to different sentences, so that they cannot be considered as belonging to the patterns we are interested in. Cases of homography are very numerous on Internet pages in Romanian: besides homonyms such as *mare* “big” and *mare* “sea”, the lack of diacritics increases in a large extent the number of homographs: for instance, *fata* may have 8 different values.

In order to solve these problems, we first introduced diacritics in these snippets using DIAC⁺ (Tufiş and Ceaşu, 2008) and then we segmented, tokenized and PoS-tagged the snippets using the TTL module (Ion, 2007). We preserved only the snippets containing NPs.

As the geographical instances we are interested in are correctly written with capital letters, we further selected only those snippets in which the seed is followed by an NP with capital initial letter.

7 Results

We manually inspected the snippets containing our seeds and, although the websites in Romanian are quite numerous, we notice that most of our seeds are not frequent. Moreover, most of the patterns considered in the experiment have a

good precision. The most useful one is *NP precum NP*, which is both frequent and very productive.

Pattern	Occurrences	Precision (%)
NP chiar și NP (“even”)	19	57.89
NP de obicei NP (“usually”)	1	100
NP în special NP (“especially”)	112	85.71
NP precum NP (“such as”)	518	95.37
NP cum ar fi NP (“such as”)	7	100

Table 3. Seeds evaluation.

In Table 4 we present the list of literals found on the web occurring in the seeds we created. For each of them we give the number of instances (NEs) found on the web that are already in the RoWN and the number of the instances (NEs) that are not in RoWN, but can be introduced (and defined by a lexicographer).

Literal	NEs already in RoWN	NEs not in RoWN
stat “state”	23	6
țară europeană “European country”	12	2
țară balcanică “Balkan country”	4	4
țară africană “African country”	15	1
țară asiatică “Asian country”	8	1
arhipelag “archipelago”	0	1
câmpie “plain”	0	2
insulă “island”	11	17
oraș “town”	36	13
fluviu “river”	3	1
capitală “capital”	11	1
cascadă “waterfall”	0	6
centru “center”	0	23
continent “continent”	5	0
imperiu “empire”	4	5
țară “country”	37	4
provincie “province”	4	14

Table 4. Seeds productivity

8 Related Work

Repositories of Named Entities are necessary for various tasks in computational linguistics. Thus, their creation can be considered a task in itself.

Toral et al. (2008) automatically extend PWN 2.1 with Named Entities using Wikipedia: the is-a hierarchy in PWN is mapped onto the Wikipedia categories; the NEs in Wikipedia are recognized and introduced in a resource called Named Entity WordNet. De Loupy et al. (2004) enrich PWN with NEs without specifying the method they used. The work that is the closest to our experiment is Mann (2002) which extracts an ontology of NEs from a news wire text using textual co-occurrence patterns, namely: common noun immediately followed by a proper noun.

Our study stands alone among these works in that it makes use of patterns reported in the literature as identified and tested for the co-occurrence of a hyponym and its hypernym at short distance in texts, with the aim of extracting instances (more exactly NEs).

9 Conclusions and Further Work

Our paper presents a new method of extracting NEs from texts, using patterns originally identified and tested for hyponym-hypernym co-occurrence in corpora. Our main aim was to test the validity of the hyponymic patterns for the instance-class relation. The second aim of this research was to enrich the RoWN with instances that could be further used for various tasks undertaken by our team. Thus, we aim at continuing the experiment with NEs from other domains. Afterwards, remarks can be made on the productivity of hyponymic patterns in various domains and on the adaptation of the methodology to each domain (e.g., in chemistry, chemical elements are considered instances, but are not normally capitalized). From such experiments we can draw conclusions about the relevance of such patterns for the hyponymy and instance-class relations, about their similarity in the way they are “lexicalized” in corpora.

Acknowledgements

The work reported here is funded by the SIR-RESDEC project, financed by the Ministry of Education, Research and Innovation under the grant no 11-007.

References

- Verginica Barbu Mititelu. 2008. Hyponymy Patterns. *Text, Speech and Dialogue, 11th International Conference, TSD 2008*:37-44.
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of COLING 2004*:101-108.
- Andrea Esuli and Fabrizio Sebastiani. 2006. Senti-WordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC 2006*: 417-422.
- Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Radu Ion and Dan Tufiş. 2007. Meaning Affinity Models. *Proceedings of the Fourth International Workshop on Semantic Evaluations*:282-287.
- Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, and Dan Tufiş. 2008. RACAI’s QA System at the Romanian-Romanian Multiple Language Question Answering (QA@CLEF2008) Main Task. *Working Notes for CLEF 2008 Workshop*.
- Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, Elena Irimia, Verginica Barbu Mititelu. 2009. A Trainable Multi-factored QA System. *Working Notes for CLEF 2009 Workshop*.
- Claude de Loupy, Eric Crestan, Elise Lemaire. 2004. Proper Nouns Thesaurus for Document Retrieval and Question Answering. *Atelier Question-Réponse, Traitement Automatique des Langues Naturelles (TALN)*.
- Gideon S. Mann. 2002. *Fine-Grained Proper Noun Ontologies for Question Answering*.
- Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*.
- Georgiana Puşcaşu, Adrian Iftene, Ionuţ Pistol, Diana Trandabăţ, Dan Tufiş, Alexandru Ceauşu, Dan Ştefănescu, Radu Ion, Constantin Oraşan, Iustin Dornescu, Alex Moruz, and Dan Cristea. Developing a Question Answering System for the Romanian-English Track at CLEF 2006. *Working Notes for the CLEF 2006 Workshop*.
- Jörg Tiedemann. 2009. News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing: Selected Papers from RANLP 2007*. John Benjamins:237-248.

- Antonio Toral, Rafael Munoz, Monica Monachini. 2008. Named Entity WordNet. *Proceedings of the 6th International Language Resources and Evaluation Conference*.
- Dan Tufiş. 2009. Paradigmatic Morphology and Subjectivity Mark-up in the RO-WordNet Lexical Ontology. In H.N. Teodorescu, Junzo Watada, and L. Jain. *Intelligent Systems and Technologies - Methods and Applications. Studies in Computational Intelligence*: 161-179.
- Dan Tufiş, Ana-Maria Barbu, Radu Ion. 2003. TREQ-AL: A Word-Alignment System with Limited Language Resources. *Proceedings of the HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*:36-39.
- Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology*, 7(2-3):9-44.
- Dan Tufiş, Radu Ion, Alexandru Ceaşu, and Dan Ştefănescu. 2006a. Improved Lexical Alignment by Combining Multiple Reified Alignments. *Proceedings of the 11th Conference on the European Chapter of the Association for Computational Linguistics*: 153-160.
- Dan Tufiş, Verginica Barbu Mititelu, Alexandru Ceaşu, Luigi Bozianu, Cătălin Mihăilă, Margareta Manu Magda. 2006b. Noi dezvoltări ale wordnet-ului românesc. In Corina Forăscu, Dan Tufiş, Dan Cristea (eds.), *Resurse Lingvistice și Instrumente pentru prelucrarea Limbii Române*: 17-22.
- Dan Tufiş and Alexandru Ceaşu. 2008. DIAC⁺: A Professional Diacritics Recovering System. *Proceedings of the 6th Language Resources and Evaluation Conference*.