Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

Increasing the Effectiveness of the Romanian Wordnet in NLP Applications*

Verginica Barbu Mititelu

Abstract

The Romanian wordnet is a semantic network under ceaseless enrichment and improvement. Its use in various applications throughout time highlighted the need for further development. In this paper we focus on a question answering scenario. We show how adding derivational relations between the literals already present in the network could help increase the effectiveness of using the Romanian wordnet in such an application. We describe the steps we took in the process of identifying, validating and adding derivational relations in our network and then simulate a question answering situation using RoWikipedia as corpus.

Keywords: wordnet, Romanian, derivational relations, question answering, lexical chains.

1 Introduction

Applications in the Natural Language Processing (NLP) domain need quality language resources for attaining good results. These resources can be lexicons, dictionaries, thesauri, grammars, etc. In this article we focus on the knowledge about words and their meanings, on the way it is represented so that to facilitate its effective use in NLP applications.

Among the various formalisms available for representing lexical knowledge, semantic networks are the most widely known and used. Furthermore, a wordnet is the most popular kind of semantic network.

*This work was supported by the Sectorial Operational Program Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HRD/89/1.5/S/59758.

It only contains nouns, verbs, adjectives and adverbs, as they make up the lexical component of a language; prepositions and conjunctions belong to the syntactic component [13], fulfilling a relational function.

In this net words are organized according to psycholinguistic principles, by means of semantic relations, many of them specific to certain parts of speech. Thus, hyponymy and meronymy are specific to nouns; hyponymy, troponymy, lexical entailment and cause are proper for verbs; descriptive adjectives are organized in clusters based on their similarity of meaning; relational adjectives are linked to the corresponding nouns, while adverbs are linked to the respective adjectives.

The first such language resource created was the Princeton WordNet (PWN henceforth) [14, 4]. Since 1985 it has been under quantitative and qualitative improvement. It served as a model for similar resources for tens of other languages. In 1996, within the EuroWordNet project [19], semantic networks started being developed for 8 European languages (Dutch, Spanish, Italian, English, French, German, Czech, and Estonian) after the model of PWN. In 2001, within the BalkaNet project [18] wordnets for Bulgarian, Czech, Greek, Romanian, Serbian and Turkish began being created (or continued being developed in the case of Czech). Nowadays there are well beyond 60 languages for which such a resource was created or is being created (a list is available here: `http://globalwordnet.org/gwa/wordnet_table.html`). The developers and users constitute a very active community, holding their conference (Global WordNet Conference) every two years in various locations around the world, Asia being a frequent host of the meetings. Moreover, all major conferences in the field of Natural Language Processing and Computational Linguistics accept papers on wordnets.

Given the success such resources have among researchers, linguistically, one can notice that the term "WordNet" has become a class name, so a common noun, and is used in the form "wordnet" to refer to any semantic network realized after the model of PWN [5].

The Romanian wordnet (RoWN henceforth) has been being developed since 2001. During BalkaNet, a common team from the Romanian Academy Research Institute for Artificial Intelligence and from the Faculty of Informatics of the "Al. I. Cuza" University of Iaşi worked for

developing a core of 18000 synsets, conceptually aligned to PWN and through it to the synsets of all the BalkaNet wordnets. The concepts considered highly relevant for the Balkan languages [18] were identified and implemented first, then a set of concepts specific to the Balkan area. After the BalkaNet project ended, the Romanian Academy Research Institute for Artificial Intelligence undertook the task of maintaining and further developing the RoWN.

We selected the concepts to be further implemented in our network so that they served other tasks that we accomplished throughout time. Thus, we aimed at a complete coverage of the 1984 corpus (`http://nl.ijs.si/ME/Vault/CD/docs/1984.html`), of the newspaper articles corpus NAACL2003 (possible to be searched for at `http://ws.racai.ro:9191`), of the Acquis Communautaire corpus (`http://ipsc.jrc.ec.europa.eu/index.php?id=198`) and of the Eurovoc thesaurus (`http://eurovoc.europa.eu/`), as much as possible from the Wikipedia lexical stock, and the verbs in VerbNet (`http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`).

We continued to follow the methodology established during the BalkaNet project, following the expand model [17]. Two basic development principles have always been followed: the Hierarchy Preservation Principle (according to which the hierarchical structure of the concepts in a wordnet is the same irrespective of the natural language for which the wordnet is developed) and the Conceptual Density Principle (which ensures that once a concept is selected to be implemented, all its ancestors up to the unique beginners are also selected, thus preventing the existence of dangling nodes) [18].

2 NLP Applications and RoWN

The ceaseless development of RoWN is (also) justified by its use in various applications implemented in our Institute. We enumerate below these applications and the way RoWN served their aims.

- Word Sense Disambiguation (WSD):

  - In a monolingual context [8]: lexical chains between the various senses of the words in the sentence are looked for in the RoWN and, when found, their length is calculated by counting the number of nodes and of edges crossed to get from one end of the chain to the other: the shorter the lexical chain, the smaller the length, so the more related the linked words senses; obviously, the lexical chain is shorter when there are more relations in the network;
  - In a multilingual context [6]: conceptually aligned wordnets for various languages permit disambiguation of homographs in one language due to their translation by different words in (an)other language(s); the results in this case were reported as better than those of WSD in a monolingual context.

- Question Answering (QA) [7]:
  - In order to automatically find the answer to a user's question formulated in natural language, the system relies only on the words introduced by the user. However, these are not always the best chosen ones. (Imagine the trivial case of non-native speakers of a language looking for information in that respective language.) That is why, it can be necessary to use also synonyms, hypernyms, hyponyms, troponyms or derived words from the ones introduced by the user. Due to its organization, a wordnet can offer access to these words for expanding the user's query, so that the sentences containing the answer could be found more easily and more reliably.
  - For ordering the answers found by the system according to their relevance in respect to the user's question, it is necessary to find a semantic similarity score between the words introduced by the user and the words occurring in the text (as they may not be the same); for calculating this score the length of the lexical chains between the respective words in the wordnet is considered. The shorter the chain, the more similar the question and the found text, so the higher the probability for it to be the answer to the user's question.

In a QA task in a multilingual context (i.e., the user asks a question in one language and needs to find the answer in texts written in a different language), conceptually aligned wordnets prove their usefulness for the cross-lingual equivalence of terms (see a detailed description in [1]).

- Machine translation: conceptually aligned wordnets for more languages are a source of equivalent words and (simple and multiword) terms useful for feeding a translation table.

3 Adding Value to the RoWN by Marking Derivational Relations

RoWN has been developed by following the expand method and obeying the Hierarchy Preservation Principle (and the Conceptual Density Principle, see above). Thus, the semantic relations in PWN have been transferred into RoWN and organize its content, too. At the moment, the distribution of synsets and literals in RoWN is indicated in Table 1.

Table 1. Statistics about RoWN – synsets and literals

Part of Speech	Synsets	Literals	Unique Literals	Non-lexicalized
Nouns	41063	56532	52009	1839
Verbs	10397	16484	14210	759
Adjectives	4822	8203	7407	79
Adverbs	3066	4019	3248	110
TOTAL	59348	85238	75656	2787

As far as semantic relations are concerned, their occurrence in our RoWN is presented in Table 2.

With the exception of "attribute" relation, all the others enumerated in Table 2 link synsets with literals of the same part of speech. A path between two words of a different part of speech, about which any speaker would say they are related, although not impossible to

Table 2. Relations in RoWN

Relation	Number
hypo/hyperonymy	48316
instance_hypo/hyperonymy	3889
antonym	4131
similar_to	4838
verb_group	1530
member_holonym	2047
part_holonym	5573
substance_holonym	410
also_see	1333
attribute	958
cause	196
entailment	371

find, would be too long, thus providing wrong information about the similarity between those words.

Besides semantic relation, PWN also contains lexical relations, which are established between literals, unlike semantic ones which hold between synsets. Lexical relations are synonymy, antonymy, derivational relations. Involving literals, they are language specific, so cannot be transferred cross-lingually. It is worth noticing in Table 2 that antonymy, which is a lexical relation in PWN, is represented as a semantic one in RoWN. The conceptual opposition between the synsets containing the antonymic pair is more useful in various applications than the mere antonymy between two literals, that is why we extended the antonymy relation from PWN at the synsets level in RoWN.

PWN also contains derivational relations. Although many of them have a correspondent in Romanian, they cannot be automatically transferred into RoWN. Such a strategy of enriching wordnets with derivational relations does exist in the wordnet community [10, 11, 12]. However, we preferred to find a language internal strategy for identifying

derivationally related words in our language and for marking them in RoWN (others who report similar attempts are [3, 15, 16, 9]). Examples of cases when there is a derivational relation in PWN but no corresponding one in RoWN between literals lexicalizing the same concepts include: *prick* - *pricker* (Romanian: *înţepa* - *sulă*), *pacify* - *pacifier* (Romanian: *împăca* - *suzetă*), *dip* - *dipper* (Romanian: *afunda* - *polonic*), etc.

In order to mark such relations in our RoWN, we followed the steps below:

1. Find possible pairs of root-derived words among the (31872) simple literals in RoWN using a list of (492) Romanian affixes and then validate the pairs. We searched for pairs of literals (literal_1 and literal_2) such that literal_1 +/− affix(es) = literal_2. The "+" version covers progressive derivation, while the "−" version covers backformation. We allow for at most 2 affixes, but of different types. The results are in Table 3.

Table 3. Derived words in RoWN

Derivation type	Derived words	Percent
Prefixation	2862	17.43
Suffixation	13556	82.57
TOTAL	16418	

We subject the found pairs to an automatic validation and then to a manual one. For the former, we relied on the information about the part of speech of the words to which affixes can attach and of the words they help create. For example, the suffix *-a* can be attached to nouns or to adjectives to create verbs.

Afterwards we proceeded to a manual validation of the whole number of pairs. The results are presented in Table 4: for each type of derivation (prefixation or suffixation), from the found pairs (column 2) we present the number of those passing the automatic validation in column 3 and then of those that passed

the manual validation in column 4; the last column presents the percent of validated pairs for each derivation type.

Table 4. Evaluation of derived words from RoWN

Derivation type	Found	Automatic Validation	Manual Validation	%
Prefixation	2862	2621	1990	69.53
Suffixation	13556	8345	8452	62.35
TOTAL	16418	10966	10442	-

2. Extract (in a set) all synsets in which each member of the above validated pairs occurs; calculate the Cartesian product of the sets for a pair of literals; validate the members of the Cartesian product, thus obtaining a list of pairs of word senses between which a derivational relation was marked (notice that it is not valid at the synset level, but at the literal one). The results are in Table 5.

Table 5. Annotated pairs in RoWN

	Prefixed	Suffixed	TOTAL
Pairs subject to validation	30132	25717	55849
Validated pairs	3145	13916	17061
Percent	10.43	89.64	30.55

3. Add a semantic label for each derivational relation in the form of a semantic relation in the network between the synsets to which the literals in derivational relation belong. A statistics of these labels can be found in [2].

Marking such relations in our wordnet, we increased the number of cross-part of speech relations to a high extent, as 66% of the suffixed

words and 97% of the prefixed words have a different part of speech from their root.

4 Short Demonstration

For proving that adding derivational relations to the RoWN we increase its effectiveness in NLP applications, let us consider the QA task. Our corpus for searching answers can be RoWikipedia. One possible question of a user is "Cine a inventat motorul cu reacţie?" ("Who invented the jet engine?"). A sentence such as "Henri Coandă a inventat motorul cu reacţie." ("Henri Coandă invented the jet engine.") does not occur in RoWikipedia. However, one can find the answer in the corpus sentence "Henri Marie Coandă (n. 7 iunie 1886 - d. 25 noiembrie 1972) a fost un academician şi inginer român, pionier al aviaţiei, fizician, inventator, inventator al motorului cu reacţie şi descoperitor al efectului care îi poartă numele." ("Henri Marie Coandă (born 7 June 1886 - died 25 November 1972) was a Romanian academician and engineer, pioneer of aviation, physicist, inventor, inventor of the jet engine and discoverer of the effect bearing his name."). The only term common to both the question and the answer is "motor cu reacţie" ("jet engine"). This unique match is not enough for giving a high score to the sentence so that it should be returned to the user. However, expanding the query, the system will also search for words that are semantically related to those introduced by the user. So, one more match will be possible: between "inventat" and "inventator". In fact, the maximum number of matches is now complete, so the sentence is retained by the system.

For calculating the semantic distance or similarity between two word senses lexical chains are created, i.e., the links and nodes in the network that are crossed for getting from one node (containing one of the target word sense) into another (containing the other target word sense). The shorter the chain, the more similar the senses. For the pair "inventa" (occurring in the user's question) - "inventator" (occurring in the corpus), the lexical chain between them crossed 6 nodes and 7 relations previously:

inventator(1.1) *instance_hyponym* James_Watt(x)
James_Watt(x) *instance_hypernym* inginer(1.1)
inginer(1.1) *hyponym* inginer_software(1)
inginer_software(1) *domain_member_TOPIC* ştiinţa_calculatoarelor(x)
ştiinţa_calculatoarelor(x) *domain_TOPIC* programa(3)
programa(3) *hyponym* crea_mental(1)
crea_mental(1) *hypernym* inventa(1)

The strangeness of this example results from the intricate path from *inventator* to *inventa*, uncommon for whatever speaker of Romanian: *inventator* - *James Watt* - *inginer* "engineer" - *inginer software* "software engineer" - *ştiinţa_calculatoarelor* "computer science" - *programa* "to program" - *crea_mental* "to create by mental act" - *inventa*. Now that derivational relations are marked, there is a direct link (semantically labeled *agent*) between the two words:

inventator(1.1) *agent* inventa(1).

5 Conclusions

Derivational relations need to be marked in a wordnet due to several reasons: derived words are part of our mental lexicon (although speakers also know the rule for creating derived words) and are in semantic relations to their roots, creating micro-networks. Moreover, from a practical perspective, the more relations are marked in the wordnet, the more effective it becomes in the applications it is used in. We have proved this in a QA scenario for Romanian. A rerun of the QA algorithm working with the enriched RoWN must support our demonstration.

References

[1] V. Barbu Mititelu, Alexandru Ceauşu, Radu Ion, Elena Irimia, Dan Ştefănescu, Dan Tufiş. *Resurse lingvistice pentru un sistem de întrebare-răspuns pentru limba română*, Revista Română de Interacţiune Om-Calculator 2 (2009), pp. 1–17.

[2] V. Barbu Mititelu. *Statistics on Derivation and its Representation in the Romanian Wordnet*, Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", pp. 99–108, 2013.

[3] O. Bilgin, O. Cetinoglu, K. Oflazer. *Morphosemantic relations in and across wordnets: A study based on Turkish*, Proceedings of GWC, pp. 60–66, 2004.

[4] C. Fellbaum (Ed.). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

[5] C. Fellbaum, P. Vossen. *The Challenge of Multilingual WordNets.* Lexical Resources and Evaluation 46, pp. 313–326, 2012.

[6] R. Ion, D. Tufiş. *Multilingual versus Monolingual Word Sense Disambiguation.* International Journal of Speech Technology, vol. 12, no 2-3 (2009), pp. 113–124.

[7] Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, Elena Irimia, Verginica Barbu Mititelu. *A Trainable Multi-factored QA System.* In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Pen as, Giovanna Roda (Eds.) Multilingual Information Access Evaluation, Vol. I Text Retrieval Experiments, pp. 257–264, Lecture Notes in Computer Science, Volume 6241/2010, Springer-Verlag.

[8] R. Ion, D. Ştefănescu. *Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization.* LTC 2009 Proc. LNCS, vol. 6562 (2011), pp. 435–443.

[9] N. Kahusk, K. Kerner, K. Vider. *Enriching Estonian WordNet with Derivations and Semantic Relations.* Proceeding of the 2010 conference on Human Language Technologies – The Baltic Perspective, pp.195–200, 2010.

[10] S. Koeva. *Derivational and Morphosemantic Relations in Bulgarian Wordnet.* Intelligent Information Systems, XVI, Warsaw, Academic Publishing House, pp. 359–389, 2008.

[11] S. Koeva, C. Krstev, D. Vitas. *Morpho-semantic relations in Wordnet – a case study for two Slavic languages.* Proceedings of the Fourth Global WordNet Conference, Szeged, pp. 239–254, 2008.

[12] K. Linden, J. Niemi. *Is It Possible to Create a Very Large WordNet in 100 days? – an Evaluation*, Language Resources and Evaluation, 2013.

[13] G.A. Miller, R. Beckwith, C. Felbaum, D. Gross, K. Miller. *Five papers on WordNet*. Technical report, Cognitive Science Laboratory, Princeton University, August 1993. Revised version.

[14] G.A. Miller. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, no. 11, pp. 39–41, 1995.

[15] K. Pala, D. Hlavackova. *Derivational relations in Czech Wordnet*. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pp. 75–81, 2007.

[16] M. Piasecki, R. Ramocki, M. Maziarz. *Recognition of Polish Derivational Relations Based on Supervised Learning Scheme*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).

[17] H. Rodriguez, S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna, A. Roventini. *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*. Computers and the Humanities, 32 (2-3), pp. 117–152, 1998.

[18] D. Tufiş, D. Cristea, S. Stamou. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*. Romanian Journal on Information Science and Technology. Special Issue on BalkaNet, volume 7, pp. 9–34, 2004.

[19] P. Vossen (Ed.). *EuroWordNet: A Multilingual Database with lexical Semantic Networks.* Kluwer. Dordrecht, The Netherlands.

Verginica Barbu Mititelu Received September 30, 2013

Research Institute for Artificial Intelligence "Mihai Drăgănescu",
Romanian Academy
13, Calea 13 Septembrie, Bucureşti 050711
Phone: +40-(0)213188103
E–mail: vergi@racai.ro