Adding Morpho-semantic Relations to the Romanian Wordnet

Verginica Barbu Mititelu
Romanian Academy Research Institute for Artificial Intelligence
13, Calea 13 Septembrie, Bucharest, Romania
vergi@racai.ro

Abstract

Keeping pace with other wordnets development, we present the challenges raised by the Romanian derivational system and our methodology for identifying derived words and their stems in the Romanian Wordnet. To attain this aim we rely only on the list of literals in the wordnet and on a list of Romanian affixes; the automatically obtained pairs require automatic and manual validation, based on a few heuristics. The correct members of the pairs are linked together and the relation is associated a semantic label whenever necessary. This label is proved to have cross-language validity. The work reported here contributes to the increase of the number of relations both between literals and between synsets, especially the cross-part-of-speech links. Words belonging to the same lexical family are identified easily. The benefits of thus improving a language resource such as wordnet become self-evident. The paper also contains an overview of the current status of the Romanian wordnet and an envisaged plan for continuing the research.

Keywords: Romanian wordnet, derivational relations, morpho-semantic relations

1. Introduction

Given the importance of language resources in the development of various tasks in computational linguistics, we consider that enriching the existing Romanian Wordnet (RoWN) with new types of information is a further step towards turning it into a knowledge base useful in question answering, information retrieval, etc.

The aims of the project presented in this paper are to mark the derivational (or morphological) relations between existing words in the RoWN and to add them a semantic label that has cross-lingual validity.

2. Related work

Wordnets have been developed for various natural languages (a list of them can be found at http://www.globalwordnet.org/gwa/wordnet_table.html). Those research groups concerned with continuous enrichment and improvement of their resource have started implementing derivational relations: Princeton WordNet (Fellbaum et al., 2004), the Czech wordnet (Pala and Hlavackova, 2007), the Turkish wordnet (Bilgin et al., 2004), the Bulgarian wordnet (Koeva, 2008), the Serbian wordnet (Koeva et al., 2008), the Estonian wordnet (Kahusk et al., 2010), the Polish wordnet (Piasecki et al., 2009).

The approaches adopted are different from one project to another. There are three main lines followed. One is that of adding morphological relations between words or word senses already existing in the wordnet (Fellbaum et al., 2004). The identification of the base-derived pairs of words is done automatically; nevertheless, manual validation proves necessary, altogether with manual grouping of pairs according to the semantic relation between the members of the pairs. The semantic relations are annotated only between verb-noun pairs of literals (not of synsets), so at the word sense level. They are available as a morphosemantic database downloadable from the http://wordnet.princeton.edu/wordnet/download /standoff/.

Another approach aims at (semi-)automatically adding new synsets to the wordnet, by automatically deriving new words from the ones already in the wordnet and linking them via morpho-semantic relations to their stems (Pala and Hlavackova, 2007; Bilgin et al., 2004; Kahusk et al., 2010). The suffixes with great productivity and clear semantics are exploited to automatically create new words, both actual and possible but unused ones, thus manual validation of the pairs becomes obligatory. For Czech and Turkish a (different) set of semantic labels was created (for each) and they were used to mark the morphologically related pairs.

The other approach, adopted for Bulgarian (Koeva, 2008), Serbian (Koeva et al., 2008), Polish (Piasecki et al., 2009), is to transfer the derivational relations existing in Princeton WordNet into wordnets aligned to it: in the target language, they are marked at the synset level and a note is added when manual inspection of the transferred relation proves that it does not hold.

3. Overview of the Romanian Wordnet

The Romanian wordnet has been under development for 11 years. At present, it is aligned to Princeton WordNet version 3.0, associated with SUMO/MILO concepts (http://www.ontologyportal.org/) and labeled with DOMAINS3.0 categories (http://wndomains.fbk.eu). It contains 51986 literals with a total of 83860 senses distributed in 57895 synsets, among which 120198 relations are established. These results were obtained from work in various national and international projects throughout the last 11 years, with the latest developments within METANET4U project (http://metanet4u.eu/), in which RoWN has already been documented with appropriate metadata and delivered as an xml file. The work of adding derivational relations associated with semantic labels that will be presented below is part of the author's postdoctoral project.

RoWN synsets contain simple literals, as well as multiple-word literals. The nouns are either common or proper (named entities). Our focus here is on simple or one-word literals and we consider only common nouns, alongside with verbs, adjectives and adverbs.

4. The Challenge: The Romanian Derivational System

New words appear in a language either by borrowing from a different language with which the former establishes contacts (due to geographic vicinity, to cultural relations, to political relations, to the spread of scientific discoveries, etc.) or by various means of combining existing linguistic material. These are, mainly, compounding and derivation. The latter makes use of suffixes, prefixes, roots and stems. A root can combine with a suffix or/and a prefix to create a new, derived word: *pădurar* "forester" is created from the root *pădure* "forest" and the suffix *-ar*. Notice that the final vowel of the root cannot be found in the derived word. Another example: *împăduri* "afforest" is obtained from the root *pădure* to which the prefix *în-* (in its phonetic variant *im-*) and the suffix *-i* are added. In all these examples, the root is also the stem. However, a stem can contain, besides the root, one or more affixes (suffixes or/and prefixes). It is base for another derived word. For example, the word *reîmpăduri* "reafforest" is created from the stem *împăduri* to which the prefix *re-* was attached. Derivation can also involve substitution of affixes: the word *despăduri* "deforest" is created by replacing the prefix *in-* with *des-* in the stem *împăduri*.

The derivation process usually lengthens the word. All examples above are cases of progressive derivation. However, sometimes derivation shortens the word, by cutting off its beginning (a prefix) or its ending (a suffix). This type of derivation is called regressive. One such example is *picta* "paint" which is a backformation from either *pictor* "painter" or *pictură* "painting".

Quite often, derivation in Romanian involves vowel or/and consonant alternations: *casă* "house" + suffix *-uță* > *căsuță* "little house" (vowel mutation), *nerod* "foolish" + suffix *-ie* > *nerozie* "foolishness" (consonant mutation), *viteaz* "brave" + suffix *-ie* > *vitejie* "bravery" (vowel and consonant mutation).

Derived words are always analyzable within a language and the etymologic information in a dictionary contains the stem and the affixes. Moreover, there are borrowings that are analyzable (not only in the language of origin, but also) in the borrowing language. For instance, the word *veselie* "cheerfulness" is a Slavic borrowing, just like *vesel* "cheerful", another Slavic borrowing. However, as there are pairs such as *hărnicie* "diligence" and *harnic* "diligent" in Romanian, where the former is derived from the latter by means of the suffix *-ie*. Both *vesel* and *harnic* are adjectives describing people, while *veselie* and *hărnicie* are nouns designating human characteristics. Given the similarities, speakers of Romanian are able to analyze *veselie* as containing the adjective *vesel* and the suffix *-ie*. Nevertheless, the etymologic information in the dictionary associated with the noun *veselie* identifies it as a Slavic borrowing. In our marking of derivationally related words, we chose to link *veselie* and *vesel* by a morpho-semantic relation for the sake of consistency in the treatment of similar cases (see the meanings of the words as presented above), as our aim is to group together semantically and morphologically related words, not to turn wordnet into an etymologic dictionary.

4.1. Romanian Suffixes and Prefixes

There are many studies in Romanian linguistics describing the Romanian affixes from various perspectives. Functionally, affixes can help create new words (and in this case they are means of derivation) or inflected forms of the same word (thus, being means of inflection). As far as their structure is concerned, affixes are simple (most of them) or complex (only a few). Some are old in language, others are newer. The former were inherited from Latin or borrowed alongside with words containing them from old Slavic, from Greek, Turkish and other languages. Since the 19th century other affixes have been borrowed (at the same time with words containing them) especially from Romance languages and even from Latin. A few affixes were created in Romanian. The affixes productivity varies throughout time. Most of them are used on the whole territory of Romania; others are restricted to certain areas. Morphologically, prefixes usually do not change the part of speech of the stem to which they attach (exceptions were presented in Petic, 2011); only suffixes can change it. The semantic values of some affixes have been synthesized in a couple of books. In spite of the monographical descriptions of a part of the affixes, we still lack an exhaustive list of Romanian prefixes and suffixes.

For our experiment, we needed a rich list of affixes and we compiled it from various bibliographical sources.

Affixes	Number
Prefixes	83
Suffixes	409
TOTAL	492

Table 1. Quantitative data about Romanian affixes

The data in this table should be interpreted like this: we have a list of 83 prefixes, but 3 of them are not used in the standard language, they are restricted dialectally; as in RoWN there are no dialectal words, these three prefixes cannot be found. As far as suffixes are concerned, they are 409 when homonymy is not considered. Otherwise, if we consider the part of speech of the word created by derivation, there are 482 suffixes. As it will be obvious from the presentation below (section 5), the part of speech of the (stem word and also of the) derived word is of extreme importance to us. Moreover, the phonetic variants of the affixes (such as *i-* and *im-* for the prefix *in-*) are also useful, so that we can automatically identify words containing affixes.

4.2. Phonetic Alternations

As mentioned above, in the process of derivation the root of the word can be affected by phonetic alternations. There are 11 possible vowel mutations and 12 consonant ones. Out of them, we have implemented so far seven vowel mutations, but only for three of them we found examples. They apply in the order in which they are enumerated here: *ea*>*e* (as in *viteaz* + suffix *-ie* > *vitejie*), *oa*>*o* (as in *floare* "flower" + suffix *-ar* > *florar* "florist"), *a*>*e* (as in *masă* "table" + suffix *-ean* > *mesean* "participant at a banquet").

5. The Methodology for Identifying Derived Words and their Base

Given the literals in the synsets of the RoWN, our first aim is to find pairs of words made up of a derived word and its base. We do not distinguish here between derivation and backformation. And we decided, at least for the moment, not to deal with derivation with a suffix and a prefix at the same time; thus, we do not find pairs like *pădure* – *împăduri*.

In order to render the steps of the derivational process, we are interested in finding for each derived word its stem, not its root (when the stem is different from the root). For instance, we want to mark *reîmpădurire* "reafforestation" derived from *reîmpăduri* "reafforest" (by means of the suffix *-re*), derived, in its turn, from *împăduri* "afforest" (by means of the prefix *re-*), which, in its turn, is derived from *pădure* "forest" by the prefix *îm-* (variant of *în-*) and the suffix *-i*. So, there are direct derivational links between stems and the words derived from them, but there are also indirect derivational links, like the one between *reîmpădurire* and *pădure*, which is reconstructed from the direct derivational links. In the figure below the direct derivational links are represented as continuous lines, while the indirect ones are represented as dashed lines. Choosing this work method is most appropriate for the derivational process that takes place in steps: affixes are usually attached one after the other. We chose not to use directed links, because we aim at a similar treatment of both proper derivation and back formation.

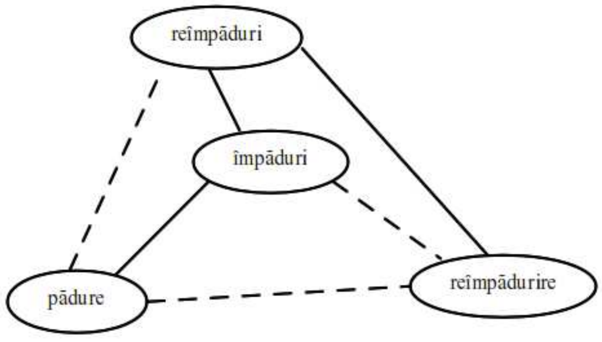

Figure 1. Direct and indirect derivational relations.

We extracted from RoWN all simple literals, irrespective of their part of speech. We did not deal with proper nouns. Given this list of literals and the list of prefixes and that of suffixes, we made combinations of one literal and either a prefix or a suffix. When the resulted form could be found in the list of literals extracted from the RoWN, we retained the pair initial literal – obtained literal as a candidate pair of a base – derived words. Adding prefixes, we obtained 2862 such pairs. Adding suffixes, we obtained 13556 pairs. The explanation comes from the fact that Romanian has a larger number of suffixes than of prefixes; suffixation is a highly productive linguistic phenomenon, unlike prefixation.

A further step was to validate these candidates. For that, we tried some automatic heuristics. For prefixes, we used a morphologic validation method: the base and the derived word must have the same part of speech. The assumption is that prefixation does not change the part of speech of the stem it attaches to. Out of the 2862 pairs only 2621 obeyed this constraint. Analyzing the eliminated pairs, we noticed that we could validate 83 of them. There are three types of examples among them: (a) 71 cases are due to RoWN incompleteness: e.g. *mulțumit* "pleased" – *nemulțumit* "displeased": these words can be adjectives, adverbs or nouns in Romanian; however, in the RoWN the former is implemented only as an adjective, while the latter as an adverb and a noun; the other values will be implemented in RoWN in the future; (b) 1 case when prefixation does change the part of speech of the base word: e.g.: *cancer* "cancer" (noun) – *anticancer* "anticancer" (adjective); this is an exception to the rule making our assumption; (c) 11 cases when the literals have a wrong part of speech tag in RoWN and require correction. Through manual inspection of the 2862 pairs of words with the same part of speech, we validated a set of 1907 base-prefixed word pairs, so almost 67%. This means that in around 33% of all cases the beginning of words is a false prefix: e.g.: *curs* "course" – *excurs* "excursus": although both words are nouns and the latter has the first two letters *ex-*, which is a prefix attachable to nouns to create another noun, the semantic condition is not fulfilled: the two words have no overlap of meaning. There are also some cases when compounding is mistaken for prefixation: *casă* "house" – *acasă* "at home" (< preposition *a* + noun *casă*).

For automatically validating the suffixed words, we exploited the morphologic information about them. Suffixes combine with words of certain parts of speech to create words with certain parts of speech. For example, the suffix *-eală* attaches to verbs to create nouns as in: *plictisi* "get bored" + suffix *-eală* > *plictiseală* "boredom".

For the suffixes occurring in our list of 13556 pairs, we established, relying on the literature dedicated to them, the parts of speech with which they combine and the part of speech of the resulting words. Exploiting this information, we numerically reduced the list to 9123 pairs. In order to establish how correct these are, we manually validated a set of (the first) 1000 pairs (that were alphabetically ordered).

In Table 2 we present the precision and recall of the methods used for finding prefixed and suffixed words. Precision is the fraction of retrieved pairs that are relevant (i.e. are pairs of base-derived words). Recall is the fraction of relevant pairs that are retrieved. While data for prefixes are for all cases found, those for suffixes are calculated for a part of 1000 pairs from the total set.

Affixes	Precision	Recall
Prefixes	70%	96%
Suffixes	71%	89%

Table 2. Precision and recall.

This means that in the case of prefixed words we are able to find almost all pairs of base-derived words (we miss only 4% of them) and we are less accurate in the case of suffixed words.

Precision is very difficult to improve: many false suffixes and prefixes cannot be spotted unless the semantics of the words is considered. Many short words (two or three letters) can be recognized within plenty other longer words, either at their beginning or at their ending, without being their roots.

However, we are more interested in increasing recall, that is in automatically finding as many pairs base-derived words as possible. Searching through wordnet for such pairs is unconceivable.

6. Marking Derivational Relations

Such morphological relations are valid only within a language and they are established at the word level, more exactly at the word sense level. Take, for example, the first sense of the verb *drive* in English. It is derivationally related with the senses 4, 8 and 11 of the noun *drive*, with the first sense of the noun *driver* and with the second sense of the noun *driving*. The sense 8 of the verb *drive* is in derivational relations with the senses 6 and 12 of the noun *drive* and with the sense 2 of the noun *driver* and sense 2 of the noun *driving*. Here is the data from the Princeton WordNet:

Sense 1
drive -- (operate or control a vehicle; "drive a car or bus"; "Can you drive this four-wheel truck?")
RELATED TO->(noun) drive#4
=> driveway, drive, private road -- (a road leading up to a private house; "they parked in the driveway")
RELATED TO->(noun) drive#11
=> drive, parkway -- (a wide scenic road planted with trees; "the riverside drive offers many exciting scenic views")
RELATED TO->(noun) drive#8
=> drive, ride -- (a journey in a vehicle (usually an automobile); "he took the family for a drive in his new car")
RELATED TO->(noun) driver#1
=> driver -- (the operator of a motor vehicle)
RELATED TO->(noun) driving#2
=> driving -- (the act of controlling and steering the movement of a vehicle or animal)

Sense 8
drive -- (push, propel, or press with force; "Drive a nail into the wall")
RELATED TO->(noun) drive#12
=> drive -- ((sports) a hard straight return (as in tennis or squash))
RELATED TO->(noun) drive#6
=> drive, driving -- (hitting a golf ball off of a tee with a driver; "he sliced his drive out of bounds")
RELATED TO->(noun) driver#2
=> driver -- (someone who drives animals that pull a vehicle)
RELATED TO->(noun) driving#2
=> driving -- (the act of controlling and steering the movement of a vehicle or animal)

Consequently, derivational relations need to be marked among literals, not at the synset level. According to wordnet terminology, these are lexical, not semantic relations. They have the following properties: (i) symmetry: if word w_1 is in derivational relation with word w_2, then w_2 is also in derivational relation with w_1; (ii) transitivity: if word w_1 is in derivational relation with word w_2 and w_2 is in derivational relation with word w_3, then w_1 is also in derivational relation with w_3 (see the indirect derivational relations represented as dashed lines in Figure 1); (iii) non-reflexivity: word w_1 is not in derivational relation with itself, which means that we do not treat conversion as a type of derivation ("zero-derivation" as it is called in various books).

7. Semantic Labeling

Usually, affixes have meanings which can be rendered in terms of semantic labels. They can be represented at the synset level.

The teams that added such labels to their wordnets worked with a different set of labels. In Princeton WordNet they are suggestive for the semantic type of the relationship between verbs and nouns. 14 labels are used: agent, material, instrument, location, by-means-of, undergoer, property, result, state, uses, destination, event, body-part, vehicle. In the Czech wordnet, they reflect the parts of speech involved in the relation rather than the semantic type of these relations: deriv-na, deriv-ger, deriv-dvrb, deriv-pos, deriv-pas, deriv-aad, deriv-an, deriv-g, deriv-ag, deriv-dem. For Turkish they were chosen so that they have a higher degree of generality: become, acquire, be-in-state, someone-with, something-with, someone-from, someone-without, something-without, pertains-to, with, reciprocal, causes, is-caused-by, cat-of, manner.

From the pairs of stem-derived words that we identified we extracted those whose members occur in only one synset each and searched for the semantic relations marked in RoWN for those synsets. We found antonymy, hypo- and hypernymy, meronymy and holonymy, pertainymy. We consider that in such cases the semantic relations are morphologically motivated and there is no need for further semantic labeling of the links.

8. Inter-lingual Transfer and Validation of Semantic Labels

Semantic labels associated with derivational relations are valid cross-lingually, even if the morphological relation is not present in all languages. For instance, in Romanian there is a morphological relation between *bucătar* "cook" and *bucătărie* "kitchen": the latter is derived from the former with the help of the suffix *-ie* (and the vowel mutation *a:ă* which is common when *a* loses stress). The semantic relations or labels involved in this case are those of agent and place. However, the same semantic relations exist for the English *cook* and *kitchen*, although they are morphologically unrelated. When wordnets are aligned, the semantic labels existing in one language can be transferred into the other language(s) or can be checked cross-lingually. Whenever discrepancies occur, they signal a mistake in annotation.

Within METANET4U project we experimented with the transfer of the semantic labels from the standoff file of Princeton WordNet. We went through 3407 pairs of synsets and for 1211 of them we found that they have equivalent morphologically related translations in Romanian. For instance, in English the verb *weed* in its second meaning "clear of weeds" is derivationally related with the first meaning of the noun *weeder* "a farmhand hired to remove weeds". Their equivalents in Romanian, *plivi* and *plivitor*, respectively, are also derivationally related and in both languages the noun expresses an agent. The other pairs (2196) either are not implemented in RoWN or the literals implementing them are not derivationally related. Consider the English pair: the verb *dry* in its second sense "become dry or drier" and the noun *drier* in its first meaning "a substance that promotes drying". Their respective equivalents in Romanian are *usca* and *sicativ*, which are morphologically unrelated.

9. Future work

The very next step in our work is to semantically annotate the pairs of base-derived words that occur in two or more synsets. For each case, we have to establish when the derivational relation holds and what semantic label should be attached to it. We already have a list of semantic labels, but it is not final and it will be adjusted according to the various situations encountered during annotation.

In order to improve the list of derivationally related words, we will implement more alternations (both for vowels and for consonants).

One more aspect that is worth investigating is the degree to which we can deal with derivation with a prefix and a suffix at the same time.

10. Conclusion

There are three levels at which the importance of marking morpho-semantic relations are evident. First, at the monolingual level, the density of relations in a wordnet increases, between words with the same part of speech, but especially between words of different parts of speech. For example, the lexical family made up of *pădure*, *pădurar*, *pădurice* "grove", *păduros* "wooded", *împăduri*, *împădurire*, *despăduri*, *despădurire*, *reîmpăduri*, *reîmpădurire*, there are four derivational links between words of the same part of speech (i.e. *pădure – pădurar*, *pădure – pădurice*, *împăduri – despăduri*, *împăduri - reîmpăduri*), and five derivational links between words of different parts of speech (noun-verb: *pădure – împăduri*, *împădurire – împăduri*, *reîmpădurire – reîmpăduri*, *despădurire – despăduri*; noun-adjective: *pădure – păduros*).

From a theoretical linguistics perspective, we can make studies concerning the semantic aspects of affixation in Romanian.

Second, at the multilingual level, the semantic labels associated with the derivational relations are established at the synset level, so they hold among concepts and could be transferred from one wordnet into another, provided that they are aligned with each other. The more wordnets with such relations, the more numerous and interesting comparative studies can be made: one can analyze how a certain semantic relation is morphologically realized in various languages: if it has a morphologic counterpart or not, what affixes express it, etc.

Third, at the applications level, a wordnet enriched with morpho-semantic relations turns into a knowledge base useful for various tasks such as question answering, information retrieval and others.

The method described here is focused only with marking morpho-semantic relations between literals already in the RoWN. In the future we could adapt our tools for marking these relations at the moment when new synsets are implemented in Romanian.

11. Acknowledgements

The main part of the work reported here is 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. Another part is supported by the METANET4U project, financed by the European Commission, under grant agreement no. 270893.

12. References

Bilgin, O., Cetinoglu, O., Oflazer, K. (2004). Morphosemantic relations in and across wordnets: A study based on Turkish. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum, P. Vossen (Eds.), *Proceedings of GWC*, 2004.

Fellbaum, C., Osherson, A., Clark, P.E. (2007). Putting Semantics into WordNet's „Morphosemantic" Links. In *Proceedings of the 3rd Language and Technology Conference*, Poznan.

Kahusk, N., Kerner, K., Vider, K. (2010). Enriching Estonian WordNet with Derivations and Semantic

Relations. In *Proceeding of the 2010 conference on Human Language Technologies – The Baltic Perspective*, pp. 195--200.

Koeva, S. (2008). Derivational and Morphosemantic Relations in Bulgarian Wordnet, *Intelligent Information Systems*, XVI, pp. 359--369.

Koeva, S., Krstev, C., Vitas, D. (2008). Morpho-semantic Relations in Wordnet – A Case Study for two Slavic Langages. In *Proceedings of the Fourth Global WordNet Conference*, pp. 239--254.

Pala, K., Hlavackova, D. (2007). Derivational relations in Czech Wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pp. 75--81.

Petic, M. (2011). Automatizarea procesului de creare a resurselor lingvistice computaționale, PhD Thesis. Institutul de Matematică şi Informatică al AŞM.

Piasecki, M., Szpakowicz, S., Broda, B. (2009). *A Wordnet from the Ground up*. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej.