Automatic Classification of WordNet Morphosemantic Relations

Svetlozara Leseva, Maria Todorova, Tsvetana Dimitrova, Borislav Rizov,
Ivelina Stoyanova, Svetla Koeva
Department of Computational Linguistics,
Institute for Bulgarian Language, Bulgarian Academy of Sciences

Abstract

This paper presents work in progress on a machine learning method for classification of morphosemantic relations between verb and noun synsets. The training data comprises 5,584 verb–noun synset pairs from the Bulgarian WordNet, where the morphosemantic relations were automatically transferred from the Princeton WordNet morphosemantic database. The machine learning is based on 4 features (verb and noun endings and their respective semantic primes). We apply a supervised machine learning method based on a decision tree algorithm implemented in Python and NLTK. The overall performance of the method reached F_1-score of 0.936. Our future work focuses on automatic identification of morphosemantically related synsets and on improving the classification.

1 Introduction

Following the observations that for languages with rich derivational morphology wordnets can recover vast amount of semantic information (Bilgin et al., 2004; Pala and Hlaváčková, 2007; Koeva et al., 2008; Barbu Mititelu, 2013), in recent years one of the main lines of research on wordnets has been focused on deciphering semantic information from derivational morphology and encoding it in and across wordnets. This paper investigates a machine learning method for classification of morphosemantic relations already identified between verb and noun synset pairs.

The morphosemantic relations as defined within the Princeton WordNet (PWN) (Agent, Undergoer, Instrument, Event, etc.) link verb–noun pairs of synsets containing derivationally related literals (Fellbaum et al., 2009). As semantic and morphosemantic relations refer to concepts, they are universal, and such a relation must hold between the relevant concepts in any language, regardless of whether it is morphologically expressed or not.

All verb and noun synsets in the PWN have been classified into semantic primes, such as person, animal, cognition, change, etc. (Miller et al., 1990), and corresponding labels, such as noun.person, noun.animal, noun.cognition, verb.cognition, verb.change have been assigned to them. Like the morphosemantic relations, the semantic primes are language independent. Moreover, there is a very strong relationship between the semantic primes of morphosemantically related synsets and the morphosemantic relation existing between them. Additional information that may be used to classify a morphosemantic relation comes from the semantics of derivational affixes.

We use the semantic primes and the derivational affixes of Bulgarian verb-noun pairs which are derivationally and morphosemantically linked in the Bulgarian WordNet (BulNet) (Koeva, 2008) as features in a machine learning method for an automatic classification of morphosemantic relations.

2 Related Work

Morphological descriptions in general lexical-semantic resources, such as wordnets (Fellbaum, 1999), Jeux de Mots (Lafourcade and Joubert, 2013) or Wolf (Sagot and Fišer, 2008) have been very popular in recent years.

The expression of morphosemantic relations through derivational means has been investigated in the wordnets of Turkish (Bilgin et al., 2004), Czech (Pala and Hlaváčková, 2007), Polish (Piasecki et al., 2012a; Piasecki et al., 2012b), Bulgarian (Koeva, 2008; Dimitrova et al., 2014), Serbian (Koeva et al., 2008), Romanian (Barbu Mititelu, 2012), among others. The work on the generation and/or identification of derivatives in a wordnet has been applied for wordnet expansion with new relations and synsets, and/or for

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 59–64,
Hissar, Bulgaria, 10–11 September 2015.

the transfer of these relations and synsets to other wordnets (Bilgin et al., 2004; Koeva et al., 2008; Piasecki et al., 2012a).

The proposal in this paper draws also on research by Stoyanova et al. (2013) and Leseva et al. (2014), who suggest approaches to filtering morphosemantic relations assigned automatically to derivationally related synsets.

3 Linguistic Motivation

In the context of wordnets, morphosemantic relations hold between synsets containing literals that are derivationally related. In the wordnet structure these relations express knowledge additional to that conveyed by semantic relations, such as synonymy, hypernymy, etc. This paper uses the inventory of morphosemantic relations from the Princeton WordNet morphosemantic database[1] which includes 17,740 links connecting 14,877 unique synset pairs by means of morphosemantic relations.

The Princeton WordNet specifies 14 types of morphosemantic relations between verbs and nouns many of which may be related to semantic roles such as agent, instrument, location, etc., though the correspondence is not always straightforward (e.g., By-means-of). The relations are: Agent, By-means-of (inanimate Agents or Causes but also Means and possibly other relations), Instrument, Material, Body-part, Uses (intended purpose), Vehicle (means of transportation), Location, Result, State, Undergoer, Destination, Property, and Event (linking a verb to a deverbal noun denoting the same event). These relations have been assigned between pairs of verb and noun synsets containing at least one derivationally related verb–noun pair of literals. For example, the noun *teacher*:2 ('a person whose occupation is teaching') is the Agent of *teach*:2 ('impart skills or knowledge to'), the noun *machine*:4 ('any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks') is the Instrument of the verb *machine*:2 ('turn, shape, mold, or otherwise finish by machinery').

A morphosemantic relation points to two types of linguistic information: (i) a (possibly) language-dependent derivational means through which literals from the respective synsets are related, and (ii) largely language-independent semantic relation of a particular type. Currently, not all pairs of verb and noun synsets containing derivationally related literals in the PWN 3.0 have been assigned a morphosemantic relation – only 7,905 out of 11,751 noun synsets derivationally related to a verb synset and 7,962 out of 8,934 verb synsets derivationally related to a noun synset have a morphosemantic relation. Moreover, in the cases where derivation is used along with other types of word formation (e.g., compounding), the synsets are not related via a derivational relation, e.g., *bookbinder*:1 'a worker whose trade is binding books' has not been linked neither derivationally, nor by means of a morphosemantic relation to *bind*:8. Finally, as the linguistic generalisations behind the morphosemantic relations have been made on the basis of the English derivational morphology, the proposed set of types and instances of relations is not exhaustive for other languages. At the same time these relations are valid in other languages, even though they might not be morphologically expressed. These considerations suggest directions for research into morphosemantic relations.

As reported by Leseva et al. (2014) for Bulgarian, the derivational patterns associated with the morphosemantic relations exhibit considerable polysemy. For example, out of 45 derivational patterns associated with the Agent relation, only 13 are monosemous. The combination of the derivational suffix and the semantic prime of the noun can be a very strong indicator for some relations. For instance, a noun with the suffix *-tel* and the semantic prime noun.person (as in *uchitel* 'teacher') is an Agent, while a noun.artifact with the suffix *-tel* (as in *dvigatel* 'engine, motor, machine') is an Instrument. Thus, even though many suffixes are ambiguous, in many cases the ambiguity can be resolved by the semantic primes. In the PWN 3.0, there are 1,142 combinations of verb–noun semantic primes within the 14,877 morphosemantically linked verb–noun synset pairs. Some of the combinations are very indicative of the morphosemantic relation, e.g., verb.contact – noun.person: Agent – 313, Undergoer – 6; verb.change – noun.substance: Result – 51; Event – 1.

4 Training Data for Machine Learning

The PWN morphosemantic relations have been transferred onto the corresponding synset pairs in

[1] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

the Bulgarian WordNet (Koeva et al., 2010). An algorithm for recognising derivationally related pairs of literals, which uses string similarity and heuristics, has been applied on the mophosemantically related synset pairs in the Bulgarian WordNet. Similarity is established if at least one of the following conditions is met: i) one of the literals is a substring of the other; ii) the two literals have a common beginning (estimated to be at least half the length of the shorter literal); iii) the two literals have a Levenshtein distance smaller than a certain threshold. Verb–noun literal pairs found to be similar have been assigned a derivational relation – prefix, suffix, or conversion (Dimitrova et al., 2014). The derivational relations have been validated manually, resulting in 6,135 relations between 5,584 unique synset pairs.

In order to improve the consistency of the dataset and to reduce noise, we have performed certain procedures on the wordnet structure: i) manual inspection and disambiguation of morphosemantic relations in case of multiple relations assigned to a synset pair; ii) validation of the consistency of the semantic primes of nouns and verbs belonging to the same natural class and the semantic primes' shift in the hypernym–hyponyms paths; iii) consistency check of the type of the assigned morphosemantic relation against the semantic primes.

4.1 Disambiguation of Morphosemantic Relations

We have identified 450 cases of multiple relations assigned between pairs of synsets, which represent 50 different combinations of two (rarely three) relations. We assume that two unique concepts are linked by a unique semantic relation, thus we keep only one relation per pair of synsets. We have distinguished several cases of multiple relation assignment, which served as a point of departure when deciding which of the relations must be preserved.

(I) One of the relations excludes the other on semantic and (frequently) syntactic grounds. Consider the assignments: <Agent, Destination>, <Agent, Undergoer>. Except in a reflexive interpretation, an entity cannot be an Agent (the doer), on the one hand, and a Destination (recipient) or an Undergoer (patient or theme), on the other. The type of relation is signalled by the synset gloss and usually by the affix. In other cases, such as <Agent, Event>, <Agent, Instrument>, the choice of relation depends on the semantic prime, e.g., a noun with the prime noun.artifact or noun.act cannot be an Agent, and vice versa–a noun.person cannot be an Instrument or an Event.

(II) One of the relations implies the other, e.g., <Instrument, Uses>, as an Instrument is used for a certain purpose. The more informative relation (in this case Instrument) has been preferred.

(III) There is no strict distinction between the relations, e.g., <Result, Event>, <Result, State>, <State, Event>, <Property, State>. In such cases, the choices are motivated on the basis of semantic information from the synsets, such as the gloss, the literals or the semantic primes. Definitions are very helpful as often they give additional information which points to the type of morphosemantic relation, e.g., 'the act of...', 'a state of...', etc. especially where the semantic prime is more specific. Certain combinations of semantic primes have been empirically established to strongly suggest the type of relation, e.g., noun.state–verb.change points to Result, noun.state–verb.state – to State. The primes noun.act and noun.event on their own have been found to be very indicative of Events. These generalisations are made after inspecting the triples noun.prime–morphosemantic relation–verb.prime.

(IV) Where other indications are lacking, we have taken into account which of the relations is more typical for a given semantic prime and/or for the synsets in the local tree (hypernyms, hyponyms, sisters).

4.2 Validation of Semantic Primes

At certain nodes in some hypernym–hyponym paths the semantic prime changes so that the hyponyms of these nodes have a different semantic prime. This may affect the homogeneity of the prime–relation correspondences. For instance, half of the Body-part relations involve the prime noun.body, and the rest – noun.animal or noun.plant. The respective nouns denote body parts or organs of animals or plants and are consistent with the definition of the prime noun.body.

We have performed a series of consistency checks on the semantic primes in chains of the type $A > B > C_1, \dots, C_n$ where A is the immediate hypernym of B, and B is the immediate hypernym of $C_1, \dots, C_n$. Five types of inconsistencies were discovered: i) the leaves (terminal hyponyms) have a different semantic prime from their immediate hypernym (the majority of the instances, 1,175 out of 1,628 for the nouns, 1,043 out of 1,607 for the verbs); ii) the non-terminal node B has a different semantic prime from A, and $C_1, \dots, C_n$ have the prime of B (382 cases for nouns, 374 for verbs); iii) some Cs have the semantic prime of A, others – of B (10 cases for nouns, 43 for verbs); iv) some Cs have the semantic prime of A, some – of B, and others – a third (different) one (43 cases for nouns, 133 for verbs); v) A and $C_1, \dots, C_n$ have the same semantic prime and B has a different one (7 cases for nouns, 14 cases for verbs).

All the cases have been manually inspected. The majority of the shifts in the semantic primes reflect specificities of the hypernym–hyponym paths, e.g., solid:18 (noun.substance) > food:3; solid food:1 (noun.food). Cases of systematic inconsistency include noun.animal or noun.plant instead of noun.body; noun.animal or noun.plant instead of noun.substance, and so forth. We have decided to keep the assigned primes and to consider assigning the primes inherited from the hypernyms in addition to the original primes.

4.3 Cross-check of Semantic Primes with Morphosemantic Relations

We have looked at the correspondences between the type of morphosemantic relations and the semantic primes of the nouns since their correlation is stronger compared to the semantic primes of the verbs. Two types of validation for consistency were carried out: i) given a noun semantic prime, which morphosemantic relations are found for the synsets of this prime and what is their frequency distribution (i.e., to what extent are they typical for a given prime); ii) given a morphosemantic relation, which noun semantic primes are found for the synsets which bear this relation and what is their frequency distribution. These checks enabled us to establish clearer criteria for the relation – semantic prime label correspondences and to reduce noise in the data. For example, the nouns linked via the relation Agent belong to 17 semantic primes, but some of them are unsuitable, such as: noun.act, e.g., *scamper:1; scramble:2; scurry:1* ('rushing about hastily in an undignified way') – an Agent of *scurry:2; scamper:2; skitter:4; scuttle:1* ('to move about or proceed hurriedly'); noun.feeling, e.g., *temper:9; mood:1; humor:7; humour:7* ('a characteristic (habitual or relatively temporary) state of feeling') – an Agent of *humor:1; humour:1* ('put into a good mood'); noun.food *dinner:1* ('the main meal of the day served in the evening or at midday') – an Agent for *dine* ('have supper; eat dinner'). The unsuitable relations have been discarded based on the nature of the relationship between the synsets, taking into account the semantic prime of the noun.

As a result of this type of validation, we have been able to reduce the nominal semantic primes associated with a morphosemantic relation, in some cases significantly: Agent from 17 to 4 (noun.person, noun.animal, noun.plant, noun.group); Instrument – from 9 to 5 (noun.artifact, noun.cognition, noun.object, noun.substance, noun.communication); Material – from 6 to 4 (noun.artifact, noun.body, noun.food, noun.substance); State – from 10 to 5 (noun.artifact, noun.body, noun.substance, noun.food); Body-part – from 4 to 3 (noun.body, noun.animal, noun.plant) but noun.body subsumes the other two; Destination is associated primarily with noun.person (i.e., Recipients), to the exception of noun.artifact (1 relation) and noun.group (2 relations); Vehicle is associated only with noun.artifact. The other 7 relations – Event, Result, Attribute, By-means-of, Uses, Location, Undergoer – show greater diversity of semantic primes and few of them could be discarded.

5 Machine Learning Task

We propose a machine learning method for automatic classification of morphosemantic relations for verb–noun synset pairs already identified as morphosemantically and derivationally related. The training is performed on a set of 5,584 labeled data instances: verb–noun synset pairs from BulNet with assigned relations (see 4).

Each data instance is represented by a combination of 4 features for the machine learning: i) verb ending (with 172 values), ii) noun ending (with 294 values), iii) verb synset semantic prime (with 15 values), and iv) noun synset semantic prime (with 25 values).

The endings are the substrings of symbols from the end of the word backwards which minimally differentiate a noun and a verb, i.e., *-sha* and *-satel* for *pisha* 'write' and *pisatel* 'writer', respectively; *--ya* and *-ach* for *gotvya* 'to cook' and *gotvach* '(a) cook', respectively; etc. The endings may include a suffix or an inflection and part of the word's base, e.g., in *pisha – pisatel*, *-sha – -satel*: *-sh-* is a root consonant and *-a* – the inflection; *-s-* is a root consonant, *-a-* is a connecting vowel, and *-tel* is the noun suffix.

This is a basic classification task which uses the set of 14 morphosemantic relations in the PWN 3.0. We apply a supervised machine learning method based on a decision tree algorithm implemented in Python and NLTK.[2] The decision tree classifier is considered suitable for the task because each pair of verb–noun synsets is assigned a single relation. Also, it performs well on large datasets in reasonable time. Moreover, we empirically confirmed that this algorithm outperformed SVM and Naive Bayes on the particular dataset.

6 Results

The evaluation is based on 10-fold cross-validation. The overall F_1 score of the morphosemantic relations classifier based on machine learning is 0.936. Table 1 shows the precision, recall and F_1 score of the method's performance across different types of morphosemantic relations.

Relation	Total	Prec	Recall	F_1
has_vehicle	3	1.000	1.000	1.000
has_agent	748	0.997	0.996	0.997
has_location	78	0.987	0.987	0.987
has_event	3,580	0.984	0.947	0.966
has_instrument	90	0.978	0.889	0.933
has_body_part_actor	5	1.000	0.833	0.917
involves_property	84	0.750	0.969	0.860
has_destination	5	1.000	0.714	0.857
has_undergoer	164	0.720	0.922	0.821
has_state	189	0.695	0.821	0.837
has_uses	123	0.691	0.850	0.771
has_result	272	0.695	0.844	0.769
by_means_of	239	0.715	0.803	0.759
has_material	10	0.000	0.000	0.000

Table 1: Evaluation of the method's performance across different morphosematic relations.

The experiment shows that the combination of the pair of semantic primes and the verb and noun endings is a relatively reliable predictor of the type of morphosemantic relation to be assigned with F_1 score ranging between 0.759 and 0.997 depending on the relation (results for relations with a low frequency in the training data are unreliable).

The analysis of the errors helped us identify the clearly defined and consistent relations (such as has_agent, has_location), as well as those that are broadly defined and thus harder to identify both by the machine learning algorithm and by human experts (has_uses, has_result, by_means_of).

7 Conclusion and Future Work

Our current research is focused on testing the performance of the method in a controlled setting on the set of derivationally related synsets in the PWN which have not been assigned a morphosemantic relation yet. In such a way we will expand the dataset and enhance the density of synset relations in BulNet. More detailed feature engineering with expert evaluation based on various features will also be tested.

The main task for our future work is to develop methods for automatic assignment of morphosemantic relations to synsets that are derivationally related but are not connected in the respective wordnet. The major challenge is given a set of derivationally related synsets in the entire wordnet, to distinguish those literal pairs (and respectively – synsets) that are semantically related from those that formally coincide.

An envisaged direction of research along these lines is to employ WordNet-based similarity measures[3] to evaluate similarity between: a) verb and noun glosses from the semantically disambiguated corpus of glosses of the Princeton WordNet;[4] b) examples of the usage of the verbs and nouns from semantically anotated corpora such as the SemCor[5] and BulSemCor (Koeva et al., 2010). The semantic similarity approach takes into account: a) the use of the verb in the noun's gloss, or vice-versa, which would mean that one is defined by means of the other; b) the presence of the verb's hypernym (on one or more steps) in the noun's gloss, or vice-versa; c) the occurrence of the verb and the noun in semantically related context; etc. Further, other components of the WordNet's structure and synset description can be applied to verify the type of the relation, including the structure of the gloss, the presence of other relations, etc.

Although our work is focused on Bulgarian and primarily uses BulNet, the results, i.e., the morphosemantic relations, are transferrable across languages and can be used to enhance wordnets for other languages with semantic content.

[2] http://www.nltk.org/_modules/nltk/classify/

[3] http://wn-similarity.sourceforge.net/

[4] http://wordnet.princeton.edu/glosstag.shtml

[5] http://www.gabormelli.com/RKB/SemCor_Corpus

References

Verginica Barbu Mititelu. 2012. Adding morpho-semantic relations to the Romanian Wordnet. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2596–2601.

Verginica Barbu Mititelu. 2013. Increasing the effectiveness of the Romanian Wordnet in NLP applications. *Computer Science Journal of Moldova*, 21(3):320–331.

Orhan Bilgin, Ozlem Cetinoglu, and Kemal Oflazer. 2004. Morphosemantic relations in and across Wordnets – a study based on Turkish. In *Proceedings of the Second Global Wordnet Conference (GWC 2004)*, pages 60–66.

Tsvetana Dimitrova, Ekaterina Tarpomanova, and Borislav Rizov. 2014. Coping with derivation in the Bulgarian WordNet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, pages 109–117.

Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting semantics into WordNet's "morphosemantic" links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics]*, volume 5603, pages 350–358.

Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.

Svetla Koeva, Cvetana Krstev, and Dusko Vitas. 2008. Morpho-semantic relations in Wordnet – a case study for two Slavic languages. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 239–254.

Svetla Koeva, Svetlozara Leseva, Ekaterina Tarpomanova, Borislav Rizov, Tsvetana Dimitrova, and Hristina Kukova. 2010. Bulgarian sense-annotated corpus – results and achievement. In *Proceedings of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages (FASSBL-7)*, pages 41–49.

Svetla Koeva. 2008. Derivational and morphosemantic relations in Bulgarian Wordnet. *Intelligent Information Systems*, pages 359–368.

Mathieu Lafourcade and Alain Joubert. 2013. Bénéfices et limites de l'acquisition lexicale dans l'expérience jeuxdemots. In *Ressources Lexicales: Contenu, construction, utilisation, valuation. Linguisticae Investigationes, Supplementa 30*, pages 187–216.

Svetlozara Leseva, Ivelina Stoyanova, Borislav Rizov, Maria Todorova, and Ekaterina Tarpomanova. 2014. Automatic semantic filtering of morphosemantic relations in WordNet. In *Proceedings of CLIB 2014, Sofia, Bulgaria*, pages 14–22.

George A. Miller, Richard Beckwith, Christiane. Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Karel Pala and Dana Hlaváčková. 2007. Derivational relations in Czech WordNet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81.

Maciej. Piasecki, Radoslaw Ramocki, and Marek Maziarz. 2012a. Automated generation of derivative relations in the Wordnet expansion perspective. In *Proceedings of the 6th Global Wordnet Conference (GWC 2012)*, pages 273–280.

Maciej Piasecki, Radoslaw Ramocki, and Pawel Minda. 2012b. Corpus-based semantic filtering in discovering derivational relations. In A. Ramsay and G. Agre, editors, *Applications – 15th International Conference, AIMSA 2012, Varna, Bulgaria, September 12-15, 2012. Proceedings. LNCS 7557*, pages 14–22. Springer.

Benoit Sagot and Darja Fišer. 2008. Building a free french WordNet from multilingual resources. In *Proceedings of the Ontolex 2008 Workshop, Marrakech, Morrocco*.

Ivelina Stoyanova, Svetla Koeva, and Svetlozara Leseva. 2013. Wordnet-based cross-language identification of semantic relations. In *Proceedings of the 4th Biennal International Workshop on Balto-Slavic Natural Language Processing*, pages 119–128.