Automatic Prediction of Morphosemantic Relations

Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova,
Tsvetana Dimitrova, Maria Todorova

Department of Computational Linguistics, Bulgarian Academy of Sciences
{svetla,zarka,iva,cvetana,maria}@dcl.bas.bg

Abstract

This paper presents a machine learning method for automatic identification and classification of morphosemantic relations (MSRs) between verb and noun synset pairs in the Bulgarian WordNet (BulNet). The core training data comprise 6,641 morphosemantically related verb–noun literal pairs from BulNet. The core dataset were preprocessed quality-wise by applying validation and reorganisation procedures. Further, the data were supplemented with negative examples of literal pairs not linked by an MSR. The designed supervised machine learning method uses the RandomTree algorithm and is implemented in Java with the Weka package. A set of experiments were performed to test various approaches to the task. Future work on improving the classifier includes adding more training data, employing more features, and fine-tuning. Apart from the language specific information about derivational processes, the proposed method is language independent.

1 Introduction

This paper investigates a machine learning method for identification and classification of morphosemantic relations (MSRs) between verb and noun synset pairs in the Bulgarian WordNet (BulNet). It is based on the MSR dataset from the Princeton WordNet (PWN) (Fellbaum et al., 2009), automatically imported to the Bulgarian WordNet (the core dataset), the PWN semantic primitives (henceforth, semantic primes) and the derivational relations (DRs) in the Bulgarian WordNet assigned automatically using a string similarity algorithm combined with heuristics (Dimitrova et al., 2014), followed by manual post-editing.

The MSRs link verb–noun pairs of synsets that contain derivationally related literals. 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. This has enabled the automatic transfer of the relations to other languages, such as Polish (Piasecki et al., 2009), Bulgarian (Koeva, 2008; Stoyanova et al., 2013; Dimitrova et al., 2014), Serbian (Koeva et al., 2008), Romanian (Barbu Mititelu, 2012; Barbu Mititelu, 2013). Other sets of MSRs have been proposed for Turkish (Bilgin et al., 2004), Czech (Pala and Hlaváčková, 2007), Estonian (Kahusk et al., 2010), Polish (Piasecki et al., 2012a; Piasecki et al., 2012b), Croatian (Šojat and Srebačić, 2014).

The study is motivated by the fact that a considerable number (53.2%, or 16,914) of the verb–noun synsets exhibiting a derivational relation (31,791 verb–noun synset pairs) in the PWN 3.0 is not labelled with an MSR. In addition, the linguistic generalisations behind the existing MSRs have been made on the basis of English derivational morphology, hence the proposed set of MSR instances may be extended based on evidence from the derivational morphology of other languages, including Bulgarian.

The present research builds on Leseva et al. (2014), where plausible MSRs were assigned by intersecting the following pairs registered in BulNet <noun literal suffix – semantic prime of the noun synset> and <noun literal suffix – MSR between the noun and a verb synset>. Then the probability for each MSR was estimated given the frequency of occurrence of the triples <MSR – noun synset semantic prime – verb synset semantic prime> in the PWN, and was used to filter out less probable MSRs.

In a follow-up paper (Leseva et al., 2015), a decision-tree based supervised machine learning method for classification of MSRs was designed, implemented and tested. In the present paper, we upgrade the previous research along the following lines – we propose a method designed to identify new synset pairs that have a high probability of being MSR related and to classify the respective MSRs; we test new sets of features combined in different ways (as described in the experiments), which gives us insights into possible extensions and improvements of the method.

Our task is three-fold: (i) to find out potential derivational verb–noun pairs in BulNet; (ii) for a given potential derivational pair, the classifier must determine whether a derivational relation exists (or there is just a formal coincidence); (iii) if a DR exists, decide what type of MSR connects the relevant synsets.

The first part of the task was implemented by identifying common substrings shared by noun–verb literal pairs and by mapping the resulting endings to the canonical suffixes. The implementation of (ii) and (iii) was performed using a machine learning classifier. The suffixes of the noun–verb derivational pairs and the semantic primes of the verb and noun synsets were used as features in the learning, while the types of MSR between these pairs of synsets were the classes in the classification task. Our research is focused on Bulgarian but the results are transferable across languages and the methodology can be used to enhance wordnets for other languages with semantic content.

2 Linguistic Motivation

2.1 Morphosemantic Relations

MSRs hold between synsets containing literals that are derivationally related and express knowledge additional to that conveyed by semantic relations, such as synonymy, hypernymy, etc. We use the inventory of MSRs from the PWN 3.0. morphosemantic database[1] which includes 17,740 links connecting 14,877 unique synset pairs. The MSRs were mapped to the equivalent Bulgarian synsets using the cross-language relation of equivalence between synsets.

The PWN specifies 14 types of MSRs between verbs and nouns: Agent, By-means-of (inanimate Agents or Causes but also Means and possibly other relations), Instrument, Material, Body-part, Uses ((intended) purpose or function), Vehicle (means of transportation), Location, Result, State, Undergoer, Destination, Property, and Event (linking a verb to its eventive nominalisation). These relations are assigned between verb–noun synset pairs containing at least one derivationally related verb–noun literal pair, e.g., *teacher*:2 ('a person whose occupation is teaching') is the Agent of *teach*:2 ('impart skills or knowledge to'). Most of the relations correspond to or are subsumed by eponymous semantic roles (Agent, Instrument, Location, Destination, Undergoer, Vehicle, Body-part, etc.).

2.2 Semantic Primes

All the verb and noun synsets in the PWN are classified into a number of language-independent semantic primes. The nouns are categorised into 25 groups, such as noun.act (acts or actions), noun.artifact (man-made objects), etc. The verbs fall into 15 groups, such as verb.body (verbs of grooming, dressing and bodily care), verb.change (verbs of size, temperature change, intensifying, etc.), as defined in the PWN lexicographer files.[2]

2.3 Derivational Relations

Derivational relations are language specific lexical relations (between pairs of literals in related synsets). A DR may signal the existence of a morphosemantic relation between the relevant synsets, which may or may not be defined explicitly in wordnet. A DR is formally expressed by means of a (combination of) morphological device(s), such as suffixation, prefixation, suffixation plus root vowel mutation, etc.

Most suffixes in Bulgarian can be associated with more than one MSR. Consider the suffix *-ach/-yach*. Its prototypical meaning is Agent, e.g., *polivach:1* (*waterer:2* – 'someone who waters plants or crops') but also denotes an instrumental meaning, e.g., *rezach:1* (*cutter:1; cutlery:2; cutting tool:1* – 'cutting implement; a tool for cutting') and other relations, such as: Vehicle – *prehvashtach:1* (*interceptor:1* – 'a fast maneuverable fighter plane designed to intercept enemy aircraft'); Body-part – *privezhdach:1* (*adductor:1* – 'a muscle that draws a body part toward the median line'); and others.

The distinction between (part of) the meanings of a suffix corresponds to a distinction in the semantic primes of the relevant noun

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

[2] https://wordnet.princeton.edu/man/lexnames.5WN.html

synsets. *Polivach:1* (Agent) has the semantic prime noun.person; *interceptor:1* (Vehicle), and *rezach:1* (Instrument) bear the semantic prime noun.artifact; *privezhdach:1* (Body-part) bears the prime noun.body. We can thus perform disambiguation or partial reduction of the number of MSRs associated with the suffix. Given a derivationally related verb–noun literal pair which has not been assigned an MSR, and a relevant suffix, we are then able to rule out the MSRs possible for that suffix but not compatible with the semantic primes of the related verb and noun synsets.

3 Linguistic Preprocessing of Training Data

We performed the following consistency procedures on the wordnet structure: (i) manual inspection and validation of MSRs in case of multiple relations assigned to a synset pair; (ii) validation of the consistency of the semantic primes in the hypernym–hyponyms paths; (iii) consistency check of the type of the assigned MSR against the semantic primes. The quality analysis and validation is performed only on the core dataset and is language independent, i.e., it concerns the wordnet structure, rather than any language data, and is transferrable across wordnets. This is a one-off task, ensuring the quality of the data used for machine learning, as well as for any future tasks based on these data.

3.1 Disambiguation of Multiple MSRs

There were 450 cases of multiple MSRs assigned between pairs of synsets, which represent 50 different combinations of two (rarely three) relations. As we assume that two unique concepts are linked by a unique semantic relation, we kept only one MSR per pair of synsets to ensure the consistency of the data. The following observations served as a main point of departure.

(I) **The relations are mutually exclusive** (24 combinations of MSRs). Consider the following assignments: <Agent, Destination>, <Agent, Undergoer>. Except in a reflexive interpretation, an entity cannot be an Agent, on the one hand, and a Destination (Recipient) or an Undergoer (Patient or Theme), on the other. The actual relation is signalled by the synset gloss and usually by the suffix, e.g., the choice of Agent over Destination for the pair *pensioner:2* (*retiree:1* – 'someone who has retired from active working') – *pensioniram se:2* (*retire:7* – 'go into retirement') was based both on the gloss and on the noun suffix *-er*. In other cases, e.g. <Agent, Event>, <Agent, Instrument>, the choice of relation depends on the semantic prime, as a noun.artifact or a 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 or overlaps with the other** (16 combinations of MSRs). Examples of such combinations are <Instrument, Uses>, <By-means-of, Instrument>, <Body-part, Uses>. The choice is based mainly on which relation is more informative rather than abstract. For example, Instrument is preferred instead of Uses as instruments are used for a certain purpose. The semantics of the suffix, e.g. *-tel* in *usilvatel:1* (*amplifier:1*) – *usilvam:7* (*amplify:1*), also plays a role in the choice of the relation (Instrument).

(III) **No strict distinction between the semantics of the relations** (10 combinations of MSRs), e.g., <Result, Event>, <Result, State>, <Result, Material>, <State, Event>, <Property, State>. The choice is motivated on the basis of semantic information from the synsets, such as the literals, the gloss, or the semantic primes. For instance, the eventive and the resultative meaning of deverbal nouns are not always distinguished as different senses. In such case, a noun.state synset would suggest the relation Result, while a noun.act or a noun.event synset points to Event. Definitions often give additional information about the type of MSR, e.g. 'the act of...', 'a state of...', etc.. By inspecting the triples <verb.prime–noun.prime–MSR>, we established prime combinations that strongly indicate the type of relation, e.g., <noun.state–verb.state> points to State; <noun.event/noun.process/noun.act–verb.change> – to Event. On their own, noun.act and noun.event point to Event, noun.person – to Agent, etc.

3.2 Validation of Semantic Primes

There are many hypernym–hyponym trees in which the semantic primes shift along the tree path. For instance, the majority of the 11,574 hypernyms with the prime noun.artifact have a hyponym classified as noun.artifact, but other prime labels are also found, such as noun.substance – for nouns denoting raw materials or synthetic substances, e.g., *pina cloth:1* ('a fine cloth made from pineapple fibers'), noun.substance, is a hyponym of *fabric:1* ('artifact made by weaving or felting or knitting or crocheting natural or synthetic fibers'), noun.artifact; etc. Moreover, some synsets are linked to two hypernyms but inherit the semantic prime of one of the two, as in: *prednisolone:1* ('a glucocorticoid (trade names Pediapred or Prelone) used to treat inflammatory conditions'), noun.substance, which is hyponym to both *glucocorticoid:1*, noun.substance, and *anti-inflammatory drug:1*, noun.artifact.

The most variation in the semantic primes of the noun synsets down a hypernym–hyponym tree is observed with: noun.state (16 other primes); noun.attribute (15); noun.group (14); etc. For example, the paths down the trees with the prime noun.group on the hypernym(s) involve noun synsets with the primes noun.person (a group of persons – for example, synsets for ethnic groups, nationalities, etc.), noun.animal (a group of animals – animal taxons, etc.), noun.plant (a group of plants – plant taxons), etc.

We analysed manually the cases where hyponyms have different semantic primes from their immediate hypernym. The primes of 33 nouns labeled as noun.Tops were changed to the prime they give name to and found predominantly in their hyponyms, e.g. *state:2* was relabelled as noun.state, *process:6; physical process:1* – as noun.process, etc. 66 hyponyms' prime labels were aligned with those of their immediate hypernym in order to reflect more precisely the semantics of the words with which they are linked. For example, *dance:2* ('move in a pattern; usually to musical accompaniment; do or perform a dance') is classified as verb.creation, its hypernym *move:14* ('move so as to change position, perform a non-translational motion') has the prime verb.motion, and *dance:2*'s hyponyms are a mix of verbs with the primes verb.creation and verb.motion. As *dance:2*'s semantics is consistent with verb.motion, the semantic prime of the verb and its hyponyms (where needed) was changed accordingly.

The majority of the shifts in the semantic primes, however, reflect specific features of the hypernym–hyponym paths – for example, the shifts between noun.substance and noun.artifact, noun.body and noun.animal or noun.plant; and so forth, especially in the cases of two hypernyms.

3.3 Cross-check of Primes and MSRs

Semantic restrictions on the combinations of semantic primes and MSRs were formulated after cross-checking their compatibility with subsequent changes either of the semantic primes of nouns and/or verbs, or of the MSR, as well as in order to reduce the number of possible combinations of <verb.prime–noun.prime–MSR> against those from the PWN 3.0. The purpose of the procedure is to ensure consistency of the training data.

The Agent is associated with persons (noun.person), social entities, e.g., organisations (noun.group), animals (noun.animal) and plants (noun.plant) that are capable of acting so as to bring about a result. Instruments are concrete man-made objects (noun.artifact), but nouns with the prime noun.communication – *debugger:1* and noun.cognition – *stemmer:3* which may function as instruments are also possible.

Inanimate causes (Fellbaum et al., 2009) – non-living (and non-volitional) entities that bring about a certain effect or result – are expressed by the MSRs Body-part, Material, Vehicle, and By-means-of. The relation Body-part may be an inanimate cause that is an inalienable part of an actor and is expressed by nouns with noun.body primes (rarely noun.animal or noun.plant). The relation Material denotes a subclass of inanimate causes – substances that may bring about a certain effect (e.g. *inhibitor:1* ('a substance that retards or stops an activity'). Beside noun.substance, noun.artifacts (synthetic substances or products) also qualify for the relation, e.g. *depilatory:2* (hair removal cosmetics). The relation Vehicle represents a subclass of artifacts (means of transportation); consequently the respective synsets have the prime noun.artifact and are generally hyponyms of the synset *conveyance:3; transport:8*. Inanimate causes whose semantics differ from that of the other three relations, are assigned the generic relation By-means-of, e.g. *geyser:2* ('a spring that discharges hot water and steam') (noun.object), etc.

The relation Event denotes processual nominalisation and involves nouns such as noun.act, noun.event, noun.phenomenon, and rules out concrete entities such as animate beings, natural (noun.object) and man-made (noun.artifact) objects, etc. The relation State denotes abstract entities such as feelings, cognition, etc. The relation Undergoer denotes entities which are affected by the event or state. The relation Result involves entities that are produced or have come to existence as a result of the event or state. The relation Property denotes various attributes and qualities. These relations involve nouns with various primes.

The relation Location denotes a concrete (natural or man-made) or an abstract location where an event takes place and therefore relates verbs with nouns with various primes – noun.location, but also noun.object, noun.plant, noun.artifact, noun.cognition, etc. The relation Destination is associated with the primes noun.person, noun.location and noun.artifact, which corresponds to two distinct interpretations of the relation – Recipient (noun.person) and Goal (noun.artifact, noun.location). The relation Uses denotes a function or purpose, e.g. *lipstick:1 – lipstick:3*. The relation allows nouns with various primes, both concrete and abstract.

We examined the combinations of noun primes and MSRs in the PWN 3.0. with a view to the semantic restrictions and in some cases MSRs were modified accordingly. For instance, some noun.body nouns were originally assigned the relation Instrument, some noun.person – Event, etc. As a result, the noun primes associated with a given MSR were reduced: Agent from 17 to 4 (person, animal, plant, group); Instrument – from 9 to 3 (artifact, communication, cognition); Material – from 6 to 2 (artifact, substance); State – from 10 to 5 (state, feeling, attribute, cognition, communication); Body-part – from 4 to 3 (body, animal, plant); Event – from 24 to 13 (act, communication, attribute, event, feeling, cognition, process, state, time, phenomenon, group, possession, relation). Result, Property, By-means-of, Uses, Location, and Undergoer are more heterogeneous and few of the semantic primes were ruled out. The relations Vehicle and Destination and the corresponding semantic primes need not be subject to any changes.

The reduction of the noun.prime–verb.prime combinations for a given MSR rules out the corresponding branches in the decision trees.

The changes made in the relations and semantic primes in these validation procedures are available at: http://dcl.bas.bg/en/wordnetMSRs.

4 Training Data for the ML Task

4.1 Core data

The core training data include examples with a confirmed MSR, after the validations procedures have been applied (see section 3) and after manual verification. The dataset comprises a total of 6,641 literal pairs in 4,016 unique synset pairs, and was compiled in two stages.

Initially, the core dataset included 6,220 instances of derivationally related verb–noun literal pairs in the BulNet verb–noun synset pairs (automatically detected and manually validated as described in Dimitrova et al. (2014)) which were assigned an MSR by automatic transfer from the PWN. We took into consideration the pairs obtained by suffixation and zero derivation.

We supplemented the core data with additional instances from BulNet extracted in the following way: (1) we identified literal pairs from BulNet which exhibited a possible DR but an MSR had not been assigned between the respective synsets; (2) after measuring the similarity of the disambiguated PWN glosses[3] for the pairs of synsets identified in step (1) using a wordnet-based measure for text similarity (Mihalcea et al., 2006), we filtered out the low similarity pairs (below threshold of 2.0); and (3) the glosses of high similarity were examined for certain structural patterns in order to determine the MSR where possible (e.g., a gloss of the type 'someone who <verb,active voice>' points to Agent, or 'instrument used for <verb>ing' – points to Instrument). As a result, 421 additional instances of morphosemantically related literal pairs were added to the core dataset.

4.2 Negative Examples Dataset

The task of determining whether an MSR holds between a given verb–noun pair is a binary classification task where the classes are *true* and *false*. To be able to train a classifier for this task, we needed a set of examples of class *false*, i.e. instances of (potentially) derivationally related verb–noun literal pairs which did not have an MSR. This can be due to various reasons: (a) one of the words has acquired an additional, usually metaphorical, meaning; (b) the similarity in the form of the noun and the verb literals is coincidental (due to historical changes in the forms, etc.) and there is no transparent DR; or (c) the relation does not fit into the pre-designed system of relations in PWN.

The negative examples were extracted automatically from BulNet and include: (i) (potentially) derivationally related verb–noun literal pairs (they

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

share a common base and the pair of endings is observed in another verb–noun literal pair with a confirmed DR) from synsets which have mutually exclusive semantic primes (i.e., not occurring among MSR pairs in PWN) and thus cannot be semantically related, e.g., verb.weather – noun.animal (total of 21,094 examples); and (ii) verb–noun literal pairs linked by a DR but not by an MSR in BulNet which formally coincide with pairs of literals that have an MSR in BulNet (over 150,000 examples). For example, the literal *gotvya* is a member of the synsets *gotvya:2* (*cook:1* – 'transform and make suitable for consumption by heating', verb.change) and *gotvya:4* (*prepare:6* – 'to prepare verbally, either for written or spoken delivery', verb.creation). The noun synset *gotvach:1* (*cook:6* – 'someone who cooks food', noun.person) derived from the verb *gotvya* bears an MSR (Agent) only to *gotvya:2*, thus the pair *gotvach:1 - gotvya:4* is extracted as a negative example.

A total of over 170,000 negative instances (verb–noun literal pairs) were extracted from BulNet. As the number and quality of the negative examples (and the number of training instances in general) affect the performance of the classifier, they usually need to be balanced against the number of positive examples and only a random selection of roughly the same number as positive data were applied in each task, equally shared between the types (i) and (ii).

4.3 Preprocessing of the Data

The Bulgarian synsets connected with MSRs from the PWN were processed using previously proposed methods and datasets. The derivationally related literal pairs found in the MS-related synsets were assigned an appropriate DR, following Dimitrova et al. (2014). The particular derivational devices were automatically established and manually validated, and the variants of the affixes (suffixes in particular) were associated with a canonical suffix form. Canonical form, as proposed in Leseva et al. (2014), is an abstract derivational pattern combining the different morphophonemic variants of suffixes.

As a first step, the word endings of each pair of verb–noun literals were identified by removing the common substring (base) shared by the two literals. In order to discard pairs that coincide in form by chance, the base was set to be at least 75% of each literal's length. Secondly, as the endings usually do not coincide with a literal's suffix (may also include part of the literal's root or stem), they were mapped to the canonical forms of the suffixes using lists of suffixes with their contextual variants. The training data contain 294 different noun endings, which were mapped to 121 canonical noun suffixes, and 172 verb endings mapped to 44 canonical verb suffixes.

In this way the number of suffix values for each MSR is reduced, while the number of examples per relation and pair of semantic primes increases, thus reducing the noise in the data that arises from the contextual suffix variants.

5 ML Method for Identification of MSRs

5.1 Features

The following features were used in the analysis of the data: (i) the canonical verb suffix; (ii) the canonical noun suffix; (iii) the semantic prime of the verb; and (iv) the semantic prime of the noun. Our data are in string format but the sets of values for both the canonical suffixes (these 121 noun and 44 verb suffixes) and the synset primes (25 semantic primes for nouns and 15 primes for verbs) are limited.

Additional features were also considered and tested such as the similarity between the glosses of the verb–noun synset pair, which was in the end disregarded due to the fact that only a limited number of instances exhibit similarity above the threshold. Instead, these examples were used to extend the training data (see section 4.1).

5.2 Implementation

The implementation of the Machine Learning is made in Java using the Weka library (Witten et al., 2011), which offers various capabilities and advanced techniques for data mining[4].

We analysed and tested various classifiers within the Weka package in order to select the best performing one suitable for the task – decision tree algorithms, Naive Bayes classifier, K* classifier, SMO (Sequential Minimal Optimisation), linear logistic regression, etc., as well as some complex classifiers applying several algorithms in a sequence. The Naive Bayes classifier was not suitable due to the data scarcity and the fact that not all combinations of feature values were covered in the data. The K* classifier relies on an entropy-based

[4] http://www.cs.waikato.ac.nz/ml/weka/

distance measure between instances and is not particularly suitable for string and nominal data. The decision tree was considered most relevant to the task. After comparing empirically several decision tree classifiers in Weka, based on the performance evaluation using 10-fold cross-validation, we selected the algorithm of RandomTree which consistently outperformed the rest. The decision tree built by the RandomTree algorithm on each node tests a given number of random features and no pruning is performed. As a baseline, we applied on the same dataset the OneR classifier which chooses one parameter best correlating with the class value to provide best prediction accuracy, and which is particularly suited for discrete data.

Three approaches were considered with a view to the method of classification. The first one uses two separate classifiers applied in a sequence – first, a binary classifier that identifies pairs of derivationally related verb–noun literals in synsets linked via an MSR, and then, a multiclass classifier that selects the type of relation. The second approach merges the above two classifiers and applies a single multiclass classifier to assign MSRs, where the set of classes includes an additional value *null* for the instances which do not have an MSR. The third method combines a set of separate binary classifiers for each of the 14 MSRs. A verb–noun pair can be assigned more than one relation, or none (in the latter case the pair is considered unrelated). The results are presented in the following section.

5.3 Experiments

Test 1. The first experiment tests the performance of the approach which first discovers whether a verb-noun pair has an MSR, and subsequently applies a multiclass classifier to assign a particular relation to the pair. The core dataset extended with 6,700 negative examples is used as training data for the binary classifier, and the classes are 'true' (there is an MSR) and 'false' (no MSR). The RandomTree classifier shows an F_1 score of 0.815 (compared to the baseline of 0.687) using 10-fold cross-validation.

The multiclass classifier is trained on the core dataset and the classes are represented by the 14 MSRs. Its F_1 score on 10-fold cross-validation is 0.842 (baseline 0.808) but varies considerably across different classes: from as high as 0.975 for Agent down to 0.333 for By-means-of (relations with less than 10 examples in the data are not considered reliable).

The F_1 score of the overall method is 0.682 since the error propagates from one phase to another. Results also show that for certain MSRs the OneR algorithm performs slightly better than the RandomTree (usually RandomTree outperforms OneR by more than 25%), which suggests that a more complex approach combining case-specific classifiers may prove more reliable.

Test 2. The second experiment tests a classifier with a list of 15 classes – the 14 MSRs and the class *null* used to label instances with no MSR. The training data include the core dataset supplemented with a limited number (6,700) of randomly selected negative examples. The results from the 10-fold cross-validation show F_1 score of 0.769 (baseline 0.654), which is significantly better than the results in Test 1. The performance also varies across relations: the highest rate is for true negatives (0.811), State (0.809), Agent (0.788), etc. In this case the RandomTree classifier significantly outperforms the baseline for all relations.

The experiment raises the question whether the negative data should be selected at random, or the training data should conform to certain selection criteria aiming at representativeness of the patterns and varieties in terms of feature values and combinations between them. Tests in this direction might be considered in the future.

Test 3. The third test examines the performance of a complex classifier combining a set of separate binary classifiers for each type of relation between a noun and a verb: there is a binary classifier (true/false) for Agent, another for Undergoer, etc. This method allows assignment of more than one relation to a given pair. In this way we can observe when uncertainty or ambiguity occurs and look for ways to tackle it. When no relation is assigned, the pair is considered unrelated. The core dataset was applied for the training of the model. In this case, for each MSR, the subset of this relation's instances constitutes the positive dataset, and the subset of instances of other relations serves as a set of negative examples.

If we look for exact matches, the results are lower: F_1 score varies from 0.81 (Agent, Event) down to $0.30 - 0.35$ (Result, By-means-of, etc.). But since in this method more than one MSR can be assigned, we can evaluate whether the correct relation is in the set of assigned relations.

Test	Baseline (OneR)	Random Tree
Test 1		
MSR true-false	0.687	0.815
Type of MSR	0.808	0.842
Overall	0.498	0.682
Test 2	0.654	0.769
Test 3		
Exact MSR	0.653	0.713
MSR in set	0.699	0.746
Reclassify *null*	0.710	0.781

Table 1: Evaluation results: F_1 score on the 10-fold cross-validation in Tests 1-3.

The method was also tested on a dataset of 300 new examples having a DR or formally coinciding with a DR, independently extracted from BulNet (not used in the training data), with their class (or lack of an MSR) manually verified. The test on an independent set is aimed at confirming that the training data are not biased and the method performs well on unknown data. Using the complex classifier, we obtained the following results: (i) exact matches are 64.00%, (ii) in another 3.33% the real class is contained in the set of guessed relations, (iii) 28.33% of the test instances are labelled as *null* while in fact they have an MSR, and (iv) the remaining 4.33% comprise incorrectly assigned relations.

The large amount of instances incorrectly labelled as *null* (28.33%) points to the need to either introduce more features to fine-tune the classifier, or to apply an additional classifier on these data using a different method, and merge results. We ran an additional classifier on all data labelled by the first classifier as *null*, using only the noun semantic prime as a feature in order to assign the most probable relation according to the semantic prime of the noun. In this case the precision increased to 78.13% by taking the most frequent relation associated with each noun prime. However, in this case we assign an MSR to all test instances, thus mislabel true negatives correctly recognised by the first classifier.

5.4 Follow-up

In further tests we experimented with variations in the data, i.e., addition of new training data instances exhibiting specific features. To this end, we assigned a second semantic prime to the synsets which either have two hypernyms (with two different semantic primes) and inherit the prime of only one of the two, or have a hypernym with another, different semantic prime which does not clash with the semantic prime of the hyponym – see the observations in 3.2. The purpose was to test whether the inherited semantic prime impacts the result. For instance, the assignment of a second prime noun.substance to synsets denoting synthetic substances or raw materials (noun.artifact) is expected to make the data more consistent as these noun.artifact synsets are more alike substances as regards the choice between certain relations, e.g., Material and Instrument. At present this shows only an insignificant increase in precision due to the small amount of data affected. However, the increase of training data in the future can potentially yield more significant improvement.

The observations on the constructed decision trees also show that the features are insufficient to fully distinguish between different MSRs as the tree structures are too shallow to achieve better results. By introducing more features, we can also test the RandomForest classification method which requires more features in order to construct a properly sized forest of RandomTree classifiers and usually outperforms the singular RandomTree method. If several learning schemes are available, it may be advantageous not to choose the best-performing one for a dataset but to use all of them and merge the results.

6 Conclusion and Future Work

Our future work will be focused on the enhancement of the method by exploring at least two mutually related directions: (i) automatic harvesting of more labelled data from other wordnets; (ii) incorporation of new features for classification and assignment of relations including heuristics derived from the WordNet structure.

Alongside the introduction of new features, it is necessary to develop techniques for reducing redundant features, as well as for correlation-based feature selection, feature ranking or principal component analysis.

We are planning to devise experiments on extended datasets with more data for English and Romanian. The multilingual data can contribute to the training with respect to the possible pairs of verb–noun primes and the relevant semantic restrictions.

While part of the information employed in this paper, such as the suffix lists and mappings from word endings to canonical suffixes, is language specific, the method proposed is language independent, including the linguistic processing of the data. Testing it for other languages is a task we envisage to implement in the future.

References

Verginica Barbu Mititelu. 2012. Adding morphosemantic 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.

Neeme Kahusk, Kadri Kerner, and Kadri Vider. 2010. Enriching Estonian WordNet with derivations and semantic relations. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 195–200, Amsterdam, The Netherlands, The Netherlands. IOS Press.

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

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.

Svetlozara Leseva, Maria Todorova, Tsvetana Dimitrova, Borislav Rizov, Ivelina Stoyanova, and Svetla Koeva. 2015. Automatic classification of wordnet morphosemantic relations. In *Proceedings of BSNLP 2015, Hissar, Bulgaria*, pages 59–64.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press.

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, Stanislaw Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground up*. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej.

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.

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.

Krešimir Šojat and Matea Srebačić. 2014. Morphosemantic relations between verbs in Croatian WordNet. In *Proceedings of the Seventh Global WordNet Conference*, pages 262–267.

Ian Witten, Eibe Frank, and Mark Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.