Automatic Prediction of Morphosemantic Relations

Svetla Koeva
Svetlozara Leseva
Ivelina Stoyanova
Tsvetana Dimitrova
Maria Todorova
Department of Computational Linguistics
Institute for Bulgarian Language, BAS

Global WordNet Conference, Bucharest, January 2016

PLAN

1. Motivation
2. Method description
    a. Datasets
    b. Method
    c. Experiments and evaluation
3. Future work

The Task

Designing and applying an ML method for automatic identification and classification of morphosemantic relations (MSRs) between verb and noun synset pairs in the Bulgarian WordNet (BulNet) based on:

- semantic primes in PWN 3.0 (Miller 1996)
- MSR data from PWN 3.0 (Fellbaum et al. 2009)
- MSR in Bulgarian – a morphologically rich language (Koeva 2008, Leseva et al. 2014)
- derivational patterns in Bulgarian (Dimitrova et al. 2014).

Morphosemantic Relations

- Adding derivational and morphosemantic relations that account for the derivational morphology in various languages;
- Cross-lingual transfer of MSRs.

Turkish (Bilgin et al., 2004); **Czech** (Pala & Hlavackova, 2007); **Bulgarian** (Koeva, 2008; Stoyanova et al., 2013; Dimitrova et al., 2014); **Serbian** (Koeva et al., 2008); **Polish** (Piasecki et al., 2009, Piasecki et al., 2012a; Piasecki et al., 2012b); **Estonian** (Kahusk et al., 2010); **Romanian** (Barbu Mititelu, 2012; Barbu Mititelu, 2013); **Croatian** (Sojat & Srebacic, 2014).

Morphosemantic Relations

Here we consider MSRs which link verb–noun pairs of synsets:

- The synset pairs contain derivationally related literals;
- There is a semantic relation between the synsets which inherits the semantics of the derivational relation between the literals.

The PWN specifies 14 types of MSRs between verbs and nouns:

Agent	*By-means-of*	*Instrument*	*Material*
Body-part	*Uses*	*Vehicle*	*Location*
Result	*State*	*Undergoer*	*Destination*
Property	*Event*		

Morphosemantic Relations

Example

teach:1 ~ teach**er**:1 '*a person whose occupation is teaching*' **Agent**

debug:1 ~ debugg**er**:1 '*a program that helps locating and correcting programming errors*' **Instrument**

arrange:5 ~ arrange**ment**:2 '*an orderly grouping; the result of arranging*' **Result**

Key points

- Derivational relations connect literals.

BUT

- Semantic and morphosemantic relations refer to concepts.
- Thus, MSRs are transferred from literals to entire synsets.
- Also, semantic relations are universal, and must hold in any language, regardless of whether they are morphologically expressed or not.

Key points

- Princeton WordNet 3.0 contains 17,740 (literal-to-literal) MSRs linking 14,476 unique synset pairs.

HOWEVER

- Part of the derivationally related verb–noun pairs of synsets in the PWN 3.0 are not labelled with an MSR.
- The MSRs are based on English derivational morphology.
- Bulgarian is a morphologically rich language with a large variety of derivational patterns and thus offers a potential source of morphosemantic information.

Objectives

The method involves:

- Identification of **potential** DRs – by identifying common substrings shared by verb–noun literal pairs and mapping the resulting endings to canonical suffixes.
- Determination of whether a derivational relation exists between a pair of **potentially** related literals or the mapping is the result of a formal coincidence;
- Classification of MSRs – determining what type of MSR links the corresponding synsets provided a DR exists.

Focus on Bulgarian

- Currently, the Bulgarian Wordnet comprises over 121,000 synsets and over 249,200 literals, linked with approx. 256,213 relations.
- Over 63,000 synsets and approx.130,000 literals have been either created or verified by experts.
- The manual validation of the automatically generated synsets includes validation, correction and supplementation of literals, glosses, examples.

Focus on Bulgarian

- There are 8,219 derivationally related (marked as such) verb–noun pairs in BulNet with no MSR assigned:

  *{podvarzvam:1} ({bind:7}) – {podvarzvach**nitsa**:1} ({bindery:1})*

- Other derivationally related pairs are not linked by a DR but are potential candidates:

  *{podvarzvam:1} ({bind:7}) – {podvarzv**ach**:1} ({bookbinder:1})*

- The results and methodology are transferable across languages

Linguistic Motivation

Semantic primes may be used to disambiguate (fully or partially) the types of MSRs for a given suffix:

Example ***-ach/-yach***

*poliv**ach**:1 (waterer:1);* prime: **noun.person** → MSR: **Agent**

*rez**ach**:1 (cutter:6);* prime: **noun.artifact** → MSR: **Instrument/Vehicle***

*prehvasht**ach**:1 (interceptor:1);* prime: **noun.artifact**→MSR: **Instr/Vehicle*** (Further restriction on Vehicle – a noun.artifact that is a hyponym of *{vehicle:1}*)

*privezhd**ach**:1 (adductor:1);* prime: **noun.body** → MSR: **Body-part**

* Partially disambiguated

Linguistic Dependencies

A couple of dependencies were taken into account:

- Verb suffix ~ Noun suffix: DR *pish**a*** - *pisa**tel*** ~ +DR
- DR ~ MSR: **-a** → **-tel** ~ Agent, Instrument, Vehicle
- Noun suffix ~ MSR: ***-ach/-yach*** ~ Agent, Instrument, Vehicle, Body-part (but not Event, or Result)
- Noun suffix ~ semantic prime: ***-tel*** ~ noun.person, noun.artifact, ...
- MSR ~ semantic prime: Agent ~ noun.person, noun.group, noun.animal

The Method

- A supervised machine learning method for MSR identification and classification.
- Based on the **RandomTree** algorithm (decision tree based on selection of features). **OneR** (frequency-based) used as baseline.
- Implemented in Java with the use of the Weka package.
- We tested different sets and combinations of features for ML.
- The proposed method, apart from the derivational processes and means, is language independent.

Machine Learning: Features

Machine Learning is based on the following features:

- Canonical noun suffix (121)
- Canonical verb suffix (44)
- Semantic prime of the noun (25)
- Semantic prime of the verb (15)

Example

zashtit**nik**:2 → **nik** (canonical) zashtit**ya**:5 → **a** (canonical)

defender — defend 'protect against a challenge or attack'

noun.person — verb.competition

Data instance: *nik, a, noun.person, verb.competition* - LABEL: Agent

Machine Learning: Training Data

The **core training dataset** comprises a total of 6,641 literal pairs in 4,016 unique synset pairs, and was compiled in two stages:

- **6,220** instances of verb–noun literal pairs with DR in BulNet, assigned an MSR by automatic transfer from the PWN.
- **421** derived by exploring gloss similarities (the Gloss Corpus).

Example

poliv**am**:1 ← possible DR → poliv**ach**:1 'someone who ...VERB...'

water:1 waterer:2 → MSR: AGENT

Gloss: '*someone who* ***waters:1*** *plants or crops*' (disambiguated PWN glosses)

Compilation and Improvement of Data

Assignment of DRs

The DRs had been assigned to the Bulgarian WordNet:

- Literals between which a derivational relation **might** exist are automatically linked using a string similarity algorithm combined with heuristics (Dimitrova et al. 2014).
- All DRs were manually verified and post-edited.

Compilation and Improvement of Data

Improvement

- Disambiguation of multiple morphosemantic relations between a unique pair of verb–noun synsets.
- Validation of semantic primes.
- Cross-check of the consistency between a semantic prime and a morphosemantic relation.

Compilation and Improvement of Data

Disambiguation of Multiple MSRs

450 cases of 2 (rarely 3) relations / 50 combinations of relations.

Semantically incompatible MSRs: Agent and Event, Agent and Undergoer, Agent and Instrument

Semantically overlapping MSRs: Instrument and Uses, Instrument and By-means-of, Instrument and Body-part

Choose the one that is consistent with the prime and is more informative.

e.g. *noun.body* is more consistent with Body-part than with Instrument.

Compilation and Improvement of Data

Validation of Semantic Primes

We analysed manually the cases where hyponyms have different semantic primes from their immediate hypernym:

- 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.
- The primes of 33 nouns labeled as **noun.Tops** were changed to the predominant prime among their hyponyms;
- 66 hyponyms' prime labels were aligned with those of their immediate hypernym;

Compilation and Improvement of Data

Validation of Semantic Primes

- many hypernym–hyponym trees in which the semantic primes shift along the tree path

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***;

- some synsets linked to two hypernyms inherit the semantic prime of one of the two

e.g., *prednisolone:1 ('a glucocorticoid used to treat inflammatory conditions')*, ***noun.substance***, which is hyponym of both *glucocorticoid:1*, ***noun.substance***, AND *antiinflammatory drug:1*, ***noun.artifact***.

Compilation and Improvement of Data

Cross-check of MSRs and Semantic Primes

- To ensure the consistency of the training data 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.
- The changes are available at: http://dcl.bas.bg/wordnetMSRs/.

Compilation and Improvement of Data

Cross-check of MSRs and Semantic Primes

The MSRs associated with a given semantic prime 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** – from10 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.
- ✓ **Vehicle** and **Destination** didn't need any changes.

Compilation and Improvement of Data

Negative Examples

Negative examples dataset was extracted automatically: pairs of noun - verb synsets with possible DR but

(1) mutually exclusive semantic primes

E.g. verb.weather – noun.animal

(2) formal coincidence of forms

E.g. *gotvya:2 (cook:1);* prime: **'verb.change'**

gotvya:4 (prepare:6); prime: **'verb.creation'** (metaphorical)

→ *gotv**ach**:1 (cook:6)*; prime: **'noun.person'** MSR only with *gotvya:2*

E.g. *lampa:1 (lamp:1)* *lamtya:1 (crave:1)* coincidence of forms

→ no MSR

Experiments

✓ **Experiment 1**: 2-step classification

(1) a binary classifier to determine whether there is an MSR, and then

(2) a multiclass classifier to assign a particular relation to the pair.

+ Relies on the fact that these are separate, independent tasks
+ May discover MSRs not covered by the 14 MSR classes
- Uses different training datasets (and different category labels) on each step
- Error propagates

F_1=0.682

Experiments

✓ **Experiment 2**: a single classifier with 15 classes – the 14 MSRs and the class *'null'* to label instances with no MSR.

\+ Reduces error compared to Experiment 1

\+ Uses one training dataset

\- Random selection of negative examples: other selection procedures may improve results

F_1=0.769

Experiments

✓ **Experiment 3**: complex classifier combining a set of separate binary classifiers for each type of relation: there is a binary classifier (‘*true*’/’ *false*’) for Agent, another for Undergoer, etc. Instances labelled as *‘false’* by all classifiers are considered without MSR.

\+ Independently trains a classifier for each MSR (more precise classifiers, e.g. for **Agent**, are not affected by less precise, e.g. **Event**)

\+ Less dependent on the amount of data (for less represented MSRs)

\+ Allows assignment of more than one MSR (overlapping MSRs)

\- Requires separate training sets

Evaluation

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: F_1 score on the 10- fold cross-validation in Experiments 1-3.

Experiment 3 results on unknown data:

(i) 64% exact matches;

(ii) 3.33% - real class is contained in the set of guessed relations;

(iii) 28.33% - labelled as null while in fact they have an MSR - further reclassified;

(iv) 4.33% - incorrectly assigned relations.

Conclusions

- A more fine-tuned method and feature design, as well as training on different sets/features in each phase, makes method more effective.
- Techniques for reducing redundant features are needed, as well as for correlation-based feature selection, feature ranking or principal component analysis.
- An additional classifier and several learning schemes may lead to objective conclusions by merging the results.

Future Work

- Enhancement of the method by:
  - ✓ exploring automatic harvesting of more labelled data from other wordnets;
  - ✓ exploring incorporation of new features for classification and assignment of relations including heuristics derived from the WordNet structure.
- Developing techniques for reducing redundant features, e.g. by correlation-based feature selection, feature ranking, etc.
- Testing the method for other languages.

References

- **Verginica Barbu Mititelu**. 2012. Adding morphosemantic relations to the Romanian Wordnet. In Proceedings of the Eighth 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), pp.109-117.
- **Christiane Fellbaum, Anne Osherson, and Peter E. Clark**. 2009. *Putting semantics into WordNet's "morphosemantic" links*. In: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics, vol. 5603, pp. 350–358.

References

- **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**. 2008. Derivational and Morphosemantic Relations in Bulgarian Wordnet. Intelligent Information Systems, pages 359–368.
- **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.
- **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**. 1996. *Natural Language Access to Intelligent Systems*. ARI Research Note 96-51 (technical report).

References

- **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.
- **Kresimir Sojat and Matea Srebacic**. 2014. Morphosemantic relations between verbs in Croatian WordNet. In Proceedings of the Seventh Global WordNet Conference, pages 262–267.

Thank you!

DCL Team

dcl@dcl.bas.bg

http://dcl.bas.bg/wordnetMSRs/