Wordnet-Based Cross-Language Identification of Semantic Relations

Ivelina Stoyanova **Svetla Koeva** **Svetlozara Leseva**

Department of Computational Linguistics, IBL, BAS, Sofia, Bulgaria

{iva,svetla,zarka}@dcl.bas.bg

Abstract

We propose a method for cross-language identification of semantic relations based on word similarity measurement and morphosemantic relations in WordNet. We transfer these relations to pairs of derivationally unrelated words and train a model for automatic classification of new instances of (morpho)semantic relations in context based on the existing ones and the general semantic classes of collocated verb and noun senses. Our experiments are based on Bulgarian-English parallel and comparable texts but the method is to a great extent language-independent and particularly suited to less-resourced languages, since it does not need parsed or semantically annotated data. The application of the method leads to an increase in the number of discovered semantic relations by 58.35% and performs relatively consistently, with a small decrease in precision between the baseline (based on morphosemantic relations identified in wordnet) – 0.774, and the extended method (based on the data obtained through machine learning) – 0.721.

1 Introduction

Natural language semantics has begun to receive due attention as many areas of natural language processing have recognized the need for addressing both the syntactic structure and the semantic representation of sentence constituents. Modelling conceptual and syntactic relationships such as semantic roles, semantic and syntactic frames, or semantic and syntactic dependencies is known as semantic role labeling – SRL (Gildea and Jurafsky, 2002), (shallow) semantic parsing (Pradhan et al., 2004), semantic role tagging (Xue and Palmer, 2004), extraction of predicate-argument structures (Moschitti and Bejan, 2004), automatic extraction of semantic relations (Swier and Stevenson, 2005), among others.

We propose a method for automatic semantic labeling based on the morphosemantic relations in Princeton WordNet (PWN). A morphosemantic relation associates a verb synset S_v and a noun synset S_n if there is a derivational relation between a literal L_v in S_v and a literal L_n in S_n. Morphosemantic relations inherit the semantics of the derivation. Consider, for instance, the morphosemantic relations *agent*, *instrument*, *location*, and *vehicle*, which link a verb to its agent (*administrate – administrator*), instrument (*collide – collider*), location (*settle – settlement*), vehicle (*bomb – bomber*).

We apply word and clause similarity measurement to parallel and comparable texts in order to perform partial word sense disambiguation and to identify candidates for labeling with semantic information. We enhance the WordNet morphosemantic relations with semantic generalizations derived from the general semantic word classes of the synsets and use this knowledge to learn and assign different types of semantic information:

- semantic relations associated with the noun collocates of a particular verb sense;
- general semantic noun classes that are eligible to collocate with a particular verb sense.

We apply this method to English and Bulgarian using PWN and the Bulgarian WordNet (BulNet). An advantage of the proposed approach is that it is able to assign semantic labels to unstructured text.

The paper is organised as follows. We outline the background against which we approach the identification of semantic relations in Section 2 where we present in brief groundbreaking and influential recent work in semantic role labeling (SRL). In Section 3 we discuss the linguistic motivation for the proposed approach. In Section 4

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 119–128,
Sofia, Bulgaria, 8-9 August 2013.

we describe the method for wordnet-based identification of semantic information and its implementation. Section 5 presents assessment of the results, followed by conclusions and an outline of directions for future research in Section 6.

2 Related Work

Many applications treat the assignment of semantic roles, semantic frames, and dependencies as a classification problem that involves the training of models on (large) manually annotated corpora, such as FrameNet text annotation (Ruppenhofer et al., 2010), the Prague Dependency Treebank (Hajic, 1998), or PropBank (Palmer et al., 2005), and the subsequent assignment of semantic labels to appropriate sentence constituents.

A number of models have been developed using the FrameNet corpus. Undoubtedly the most influential one has been Gildea and Jurafsky's machine learning method (Gildea and Jurafsky, 2002), which is based on the training of a SRL classifier on a set of lexical, morpho-syntactic, syntactic and word order features extracted from the parsed FrameNet corpus in conjunction with knowledge of the predicates, prior probabilities of various combinations of semantic roles, etc.

PropBank spurred a lot of research in SRL (Pradhan et al., 2004; Pradhan et al., 2008; Toutanova et al., 2008; Surdeanu et al., 2003; Xue and Palmer, 2004), to mention but a few. For instance, Pradhan et al. (2004) and Pradhan et al. (2008) propose SRL algorithms that augment previously developed systems, such as Gildea and Jurafsky's (2002) by replacing earlier classifiers with SVMs. Xue and Palmer (2004) train a Maximum Entropy classifier on the PropBank using linguistic features that can be directly extracted from syntactic parse trees and achieve results comparable to the best performing system at the time (Pradhan et al., 2004).

Semantic role labeling based on (large) annotated corpora need to deal with a number of issues, such as the situation specificity of semantic roles, the manual selection of annotated examples, variability in the sets of roles used across the computational resources, among others (Marquez et al., 2008). Pradhan et al. (2008) have also shown that the transfer of such models to other domains leads to substantial degradation in the results.

Some researchers employ other resources as an alternative. Swier and Stevenson (2005) describe an unsupervised SRL system that combines information from a verb lexicon – VerbNet with a simple probability model. Shi and Mihalcea (2005) propose the integration of VerbNet, WordNet and FrameNet into a knowledge base and use it in the building of a semantic parser. The system identifies the FrameNet frame best corresponding to a parsed sentence either as a direct match, or via VerbNet and/or WordNet relations.

Despite these alternatives the dominant trend has remained the corpus-based SRL, with unsupervised approaches gaining popularity as a way of overcoming the deficiencies of supervised methods (Lang and Lapata, 2011a; Lang and Lapata, 2011b), among others. Syntactic analysis has been considered a prerequisite in SRL, with full parsing winning over partial parsing, as demonstrated by the results in the CoNLL-2004 (Carreras and Marquez, 2004) and the CoNLL-2005 (Carreras and Marquez, 2005) shared tasks. Syntactic analysis and SRL have been dealt with within two general frameworks. In the "pipeline" approach the systems first perform syntactic parsing followed by SRL, while In the joint parsing approach syntactic and semantic parsing are performed together. Joint parsing of syntactic and semantic dependencies has been the focus of the CoNLL-2008 (Surdeanu et al., 2008) and CoNLL-2009 (Hajič et al., 2009) shared tasks.

To sum up, a classical SRL system takes a parsed input and assigns semantic roles on the basis of: i) a language model learnt from a pre-annotated semantically labeled corpus; ii) a frame lexicon; or iii) a combination of different resources. In the systems using annotated corpora the syntactically parsed sentences are usually semantically annotated using classifiers trained on the corpus on the basis of linguistic features derived from the parses. In the case of lexicon-based systems semantic roles are directly or indirectly assigned from the lexicon.

3 Motivation

Morphosemantic relations in WordNet denote relations between (synset members) that are similar in meaning and where one word is derived from the other by means of a morphological affix (Fellbaum et al., 2009). The authors also note that most of the morphosemantic relations connect words from different classes and go on to demonstrate that part of the noun-verb relations correspond to semantic roles. In fact, many of the noun-verb morphosemantic links in WordNet designate typical relations between a participant and a predicate, such as *agent*, *instrument*, *material*, *location*, *undergoer*, *destination*, etc.

For instance the verb literal *send* (cause to be directed or transmitted to another place) is related to the noun *sender* (someone who transmits a message) through the morphosemantic relation *agent* and to the noun *sendee* (the intended recipient of a message) through *destination*; *train* (educate for a future role or function) is connected to *trainer* (one who trains other persons or animals) through *agent* and to *trainee* (someone who is being trained) through *undergoer*. The noun members of these morphosemantic relations can function as arguments of the particular verbs and bear the respective semantic roles, i.e. *agent* for *sender* and *trainer*, *destination* for *sendee*, and *undergoer* for *trainee*.

Further, we assume that if a noun and a verb enter into the same morphosemantic relation individually, they are licensed for it and therefore, when they collocate, they enter into this relation if there is no other appropriate noun candidate for the same relation. As an example, consider the sentence: *The author used the method of cost-effectiveness analysis*. The verb *use* is linked to *user* through the morphosemantic relation *agent*. The noun *author* is connected with the verb *author* (be the author of) by means of the same relation. By virtue of the above assumption we assign the relation *agent* between *use* and *author* in this particular sentence. In such a way the morphosemantic relation identified between the derivationally related verb and noun may be inherited by synonyms, direct hypernyms, hyponyms, sister terms, etc. Thus, given a morphosemantic relation and words in the context that participate in such a relation independently of each other, we are able to discover certain types of semantic relations.

4 Method for Cross-Language Learning of Semantic Relations

The goal of the study is to identify semantic relations between a verb and collocated nouns[1] within similar clauses in Bulgarian and English (often but not necessarily translational equivalents) and to assign a semantic matrix to the verb based on collocational evidence and the WordNet hierarchy.

	Bulgarian	English
Administrative		
Politics	28,148	27,609
Economy	25,800	28,436
Health	26,912	30,721
Ecology	27,886	36,227
News		
Politics	25,016	25,010
Economy	25,010	25,127
Culture	25,319	25,355
Military	25,283	25,328
Fiction		
Adventure	25,053	29,241
Humour	30,003	26,992
Love	32,631	25,459
Fantasy	30,200	32,393
TOTAL	327,261	337,898

Table 1: Distribution of texts (in terms of number of words) in the Bulgarian-English comparable corpus applied in the study

The method is developed and tested on a Bulgarian-English comparable corpus (Table 1) which is an excerpt from the Bulgarian National Corpus (Koeva et al., 2012).

4.1 WordNet Enhancement with Morphosemantic Relations

The interest in morphosemantic relations has been motivated by the fact that they overlap to a great extent across wordnets (Bilgin et al., 2004) and thus improve the internal connectivity of the individual wordnets, as well as by the fact that the derivational subnets reflect certain cognitive structures in natural languages (Pala and Hlavackova, 2007). n approach to wordnet development based on enrichment with morphosemantic relations has been adopted for English (Fellbaum et al., 2009), as well as for a number of other languages – Turkish (Bilgin et al., 2004), Czech (Pala and Hlavackova, 2007), Bulgarian (Koeva et al., 2008), Serbian (Koeva, 2008), Polish (Piasecki et al., 2009), Romanian (Barbu Mititelu, 2012), to mention a few.

Provided there is a mapping algorithm between two or more wordnets, such as the cross-language relation of equivalence between synsets (Vossen, 2004), a morphosemantic relation between a pair of synsets in a given language can be mapped to the corresponding synsets in a different lan-

[1] Collocated here means that nouns are found within the same clause as the verb.

guage, even if the latter language does not exhibit a derivational relation between members of these particular synsets.

We automatically expand BulNet with morphosemantic relations in the following two ways:

(1) Morphosemantic relations are mapped from the morphosemantic database distributed with the PWN[2] to the corresponding Bulgarian synsets. The morphosemantic relations currently encoded in Princeton WordNet 3.0.[3] have relatively limited coverage – 14,876 verb-noun synset pairs, which involve 7,960 verb synsets and 9,704 noun synsets. The automatic transfer of morphosemantic links to BulNet resulted in the association of 5,002 verb-noun pairs involving 3,584 verb synsets and 4,938 noun synsets.

For example, the PWN synset *hammer:2* (beat with or as if with a hammer) is related to the noun synset *hammer:4* (a hand tool with a heavy rigid head and a handle; used to deliver an impulsive force by striking) through the morphosemantic relation *instrument*. We map this relation to the corresponding pair in BulNet – the verb synset *chukam:1; kova:1* and the noun synset *chuk:1*. In the particular case a derivational relation exists in Bulgarian, as well, between *chuk* and *chukam*.

(2) In general, the task of detection and classification of the identified relations includes automatic generation of derivational pairs based on knowledge of language-specific derivational patterns followed by filtering of the results through automatic and/or manual validation. Specific methods are described in more detail in the research cited at the beginning of this subsection, as well as in more recent proposals, such as the machine learning approach to generation and classification of derivational pairs made by Piasecki et al. (2012b) and Piasecki et al. (2012a), respectively.

We identify new pairs of verb-noun literals in BulNet that are potentially derivationally (and thus morphosemantically) related by means of a set of rules that describe the verb-noun and noun-verb derivational patterns in Bulgarian (we focus on patterns affecting the end of the word, thus ignoring prefixation) and assign the respective morphosemantic relations to the synsets that include the related pairs.

We identified 89 derivational noun endings (morphophonemic variants of suffixes) and 183 derivational patterns (verb ending to noun ending correspondences), and associated them with the morphosemantic relation they indicate. Only 39 of the selected derivational endings were found to be unambiguous. Moreover, many of them proved to be highly ambiguous, denoting up to 10 of the 14 morphosemantic relations. In order to disambiguate at least partially the possible morphosemantic relations associated with a particular suffix, we filtered those meanings with the general semantic classes derived from the PWN lexicographer files. The PWN synsets are organized in 45 lexicographer files based on syntactic category and general semantic word classes (26 for nouns and 15 for verbs)[4].

For instance, the Bulgarian noun suffix -*nik* is associated with the following relations *agent*, *instrument*, *location*, *undergoer*, and *event*. By virtue of the fact that the synsets denoting locations are found in the lexicographer file `noun.location`, the synset denoting agents – in `noun.person`, and the instruments – in `noun.artifact`, we were able to disambiguate the suffix at least partially.

Initially, 57,211 new derivational relations were found in BulNet. These relations were evaluated automatically on the basis of the morphosemantic relations transferred from PWN. Each triple `<verb.label, noun.label, relation>`[5] was assigned a probability based on the frequency of occurrence in the set of morphosemantic relations transferred from PWN. Those relations with a probability below 1% were filtered out. As a result 34,677 morphosemantic relations between a noun literal and a verb literal were assigned among 7,456 unique noun-verb synset pairs, which involved 2,537 verb synsets and 1,708 noun synsets.

For example the noun synset *kovach:1* (corresponding to *blacksmith:1*) is derivationally related with the verb literal *kova:1* through the suffix *-ach*, which is associated either with an *agent* or with an *instrument* relation depending on the semantics of the noun – a person or an inanimate object. In this case the meaning of the suffix is disambiguated

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

[3] http://wordnet.princeton.edu/wordnet/download/standoff/

[4] http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html

[5] The `verb.label` and `noun.label` are descriptive labels of the wordnet synsets which are listed in the respective lexicographer files.

by virtue of the fact that *kovach:1* is found in the `noun.person` lexicographer file. We link the literals *kova:1* and *kovach:1* via a derivational relation *suffix* and assign the synsets the morphosemantic relation *agent*.

4.2 Preprocessing and Clause Splitting

The preprocessing of the Bulgarian-English corpus used in the study includes sentence-splitting, tokenization, POS-tagging and lemmatization, using the Bulgarian Language Processing Chain[6] (Koeva and Genov, 2011) for the Bulgarian part and Stanford CoreNLP[7] for the English part.

The clause serves as the minimal context for the realization of verb semantics, and hence – the scope within which we carry out the cross-linguistic analysis and the assignment of relations. Clause splitting is applied using a general method based on POS tagging, lists of clause delimiters (clause linking words, multiword expressions, and punctuation) and a set of language-specific rules. We define the clause as a sequence of words between two potential clause delimiters where exactly one predicate (a simple or a complex verb form, which may be a lexical verb, an auxiliary, a copula, or a combination of a lexical verb or a copula with one or more auxiliaries) occurs. We identify the predicates in each sentence using language-specific rules for Bulgarian and English. Each clause is labeled by a clause opening and a clause end. The clause splitting algorithm marks subordinating and coordinating clause linking words and phrases and punctuation clause delimiters. If no clause boundary has been identified between two predicates, a clause boundary is inserted before the second one. The nested clauses are detected, as well.

4.3 Word-to-Word and Text-to-Text Semantic Similarity

WordNet has inspired the elaboration of metrics for word similarity and relatedness that quantify the degree to which words (concepts) are related using properties of the WordNet structure. The so-called path-length based measures rely on the length of the path between two nodes (synsets), possibly normalized. For instance, the Leacock-Chodorow metric (Leacock and Chodorow, 1998) finds the shortest path between two concepts and scales the path length by the overall depth D of the WordNet taxonomy, while Wu-Palmer (Wu and Palmer, 1994) calculates the depth of the concepts and their least common subsumer in the WordNet taxonomy.

Information content based metrics augment the path information with corpus statistics. Resnik (1995) measures the similarity of two concepts by calculating the information content (IC) of their least common subsumer (LCS). Jiang-Conrath (Jiang and Conrath, 1997) and Lin (Lin, 1998) combine the information content of the LCS with the information content of the individual concepts.

Several relatedness metrics have also been proposed, such as Hirst-St-Onge (Hirst and St-Onge, 1998), which measures semantic relatedness based on the path length and its nature (the changes of direction in the path), and the algorithms proposed by Banerjee and Pederson (2002) and Patwardhan et al. (2003), which rely on information obtained from the synsets glosses.

A number of researchers have addressed WSD based on cross-lingual semantic similarity measurement, such as the application of monolingual WSD graph-based algorithms to multilingual co-occurrence graphs based on WordNet (Silberer and Ponzetto, 2010), or of multilingual WSD algorithms based on multilingual knowledge from BabelNet (Navigli and Ponzetto, 2012).

For the purposes of the extraction of semantic relations we are interested in corresponding pairs of clauses in Bulgarian and English satisfying the following conditions: (a) the verbs in the clauses are similar (with respect to a certain similarity measure and threshold); and (b) the clauses are similar in meaning (with respect to a certain similarity measure and threshold). Similar pairs of verbs and nouns are identified on the basis of the Wu-Palmer word-to-word similarity measure (Wu and Palmer, 1994). Clause similarity is computed by means of the text similarity measurement proposed by Mihalcea et al. (2006).

Measuring semantic similarity cross-linguistically enables us to filter some of the senses of a particular word in one language since potential semantic similarity of words within similar clauses strongly suggests that these words are semantically related – translation equivalents, close synonyms, hypernyms, or hyponyms.

In the application of the method described in Section 4.4, we assign semantic relations to the el-

[6] `http://dcl.bas.bg/services/`

[7] `http://nlp.stanford.edu/software/corenlp.shtml`

ements of similar clauses in a comparable, not necessarily parallel, Bulgarian-English corpus. Moreover, we identify semantically similar rather than parallel clauses, which enables us to experiment with a greater number and diversity of contexts for the identification of semantic relations.

4.4 Method outline

We select corresponding (comparable) pairs of texts from the corpus – T_1 in Bulgarian and T_2 in English on the basis of their detailed metadata description (Koeva et al., 2012), including parameters such as style, domain and genre. For each pair of corresponding texts T_1 and T_2 we apply the following algorithm:

Step 1. We identify semantically similar pairs of verbs and consider similarity between their respective clauses – $v_1 \in \mathrm{cl}_1$ and $v_2 \in \mathrm{cl}_2$, where $\mathrm{cl}_1 \in T_1$ and $\mathrm{cl}_2 \in T_2$ and cl_1 are also semantically similar (cf. Section 4.3 for word-to-word and clause-to-clause similarity).

Step 2. We identify semantically similar pairs of collocated nouns in the bi-clause $(\mathrm{cl}_1, \mathrm{cl}_2)$ in the same way as for verbs.

Step 3. We assign morphosemantic relations to the verb and its collocated nouns using the enhanced set of relations (cf. Section 4.1) and map all matching candidates (v_1, n_1, rel) in $\mathrm{cl}_1(v_1)$ and (v_2, n_2, rel) in $\mathrm{cl}_{(}v_2)$.

Step 4. Since co-occurrence of members of a single instance of a morphosemantic relation are relatively rare, we transfer the morphosemantic relations to non-derivationally related words, provided a noun and a verb participate in the same type of morphosemantic relation independently of each other. In Example 1 the noun *director* enters into a morphosemantic relation (*agent*) with the verb *direct*, while the verb *send* enters independently into the same type of relation with the noun *sender*. Since both *director* and *send* are licensed for *agent*, we assign the relation.

Example 1.
Croatian director Zrinko Ogresta sent an invitation for the international film festival.
send, 01437254-v, verb.contact
{to cause or order to be taken directed or transmitted to another place}
director, 10015215-n, noun.person
VERB *send*: agent, NOUN *director*: agent_inv

Step 5. We hierarchize the candidates for a particular morphosemantic relation and select the most probable one based on the general semantic word classes (`verb.label` and `noun.label`) and the relations they participate in. Where two or more morphosemantic relations are assigned between a pair of words, priority is given to the relation which is most compatible with the general semantic class of the noun in the relation.

Some relations, such as *event*, are very general and therefore are not considered even if their probability is higher, provided a more meaningful relation is available. Moreover, we incorporate some syntactic and word-order dependencies. For instance, a noun which is a complement of a prepositional phrase and is thus found in the following configurations: $p(A)N$ (with any number of adjectives preceding the noun) is not licensed for the morphosemantic relation *agent* if the verb is active.

Step 6. Based on the general semantic class of the noun and/or the verb, some additional potential relations are added to the respective synsets in the model (Example 2). For example, a noun belonging to the class `noun.location` can potentially enter into a *location* relation with the verb, although the respective noun synset might not enter into this morphosemantic relation.

Example 2. Newly added relations

`verb.label`	**role**
contact	agent
motion	location

`noun.label`	**role**
person	agent_inv
location	location_inv
attribute	property_inv

Step 7. We extend the number of relations by learning from previous occurrences. Learning is performed on the News subcorpus (see Table 1), and further experiments extend the information acquired in the learning phase with data from the entire corpus.

Given a verb is licensed to enter into a particular morphosemantic relation, we assign this relation to a co-occurring verb-noun pair, even if the noun in this pair does not enter into this type of relation, provided other nouns belonging to the same general semantic class have been observed to cooccur with this verb. This assumption is generalized over all the members of the verb synset and the noun synset to which the respective verb and noun belong.

Example 3 shows how learning is applied: based on the occurrences of verbs from the same synset (ID: 00974367-v, *announce:2; declare:2*) in a morphosemantic relation of type *agent* with nouns belonging to the semantic class `noun.group` (in 60.4% of the cases), we associate the verb *announce* with the noun *Ministry* (`noun.group`) through the *agent* relation despite the fact that *Ministry* is not linked to any verb through a morphosemantic relation.

Example 3.

Learned:

Verb ID	relation	noun.label	freq
00974367-v	by-means-of	noun.artifact	5
00974367-v	by-means-of	noun.act	14
00974367-v	agent	noun.person	9
00974367-v	agent	noun.group	16

The Ministry of Defense announced on Wednesday its new plans.
announce, 00974367-v, verb.communication
{make known, make an announcement}
Ministry, 08114004-n, noun.group
VERB *announce*: agent
NOUN *Ministry*: agent_inv

At a next stage of generalization we consider only the general semantic classes of a verb and a noun which are candidates to enter in a morphosemantic relation. This step relies on the assumption that verbs from the same semantic class (e.g. perception verbs) show preference to similar semantic patterns. The learned information is in a generalized form, as presented in Example 4.

Example 4. A sample of semantic compatibility information learned from the News subcorpus.

verb.label	relation	noun.label	freq
verb.perception	undergoer	noun.person	15
verb.perception	undergoer	noun.group	3
verb.perception	state	noun.state	12
verb.perception	by-means-of	noun.state	12
verb.perception	by-means-of	noun.act	6
verb.perception	uses	noun.group	3
verb.perception	agent	noun.person	3

4.5 Implementation

We implement the word-to-word similarities with the *ws4j* package for Java[8], which is based on the original Perl package `Wordnet::Similarity` (Pedersen et al., 2007).

We use the Princeton WordNet 3.0 and access it through Java libraries such as JAWS[9] and JWI[10]. We also employ a list of morphosemantic relations available for WordNet 3.0. The access to BulNet is modeled roughly on PWN. The corresponding synsets in the two wordnets are linked by means of synset IDs.

5 Results and Evaluation

Evaluation was performed with respect to the coverage of the morphosemantic relations, the precision of the assigned relations, and the informativeness of the extracted semantic patterns. Testing was carried out on the News subcorpus (Table 1) totaling 100,628 tokens distributed in four subdomains: Politics, Economy, Culture, and Military. The corpus comprises 3,362 sentences and 7,535 clauses for Bulgarian and 3,678 sentences and 8,624 clauses for English.

	Method	# clauses	# relations
1	Baseline_0	920	$1,183$
2	Baseline	951	$1,246$
3	Learned and transferred to synsets	$1,032$	$1,353$
4	Learned and transferred to semantic classes	$1,395$	$1,973$

Table 2: Coverage of relations in the News subcorpus using the baseline method (2) and the extended method (4)

Table 2 presents: (1) the number of morphosemantic relations covered by the baseline_0 method, i.e. applying only the Princeton WordNet morphosemantic relations; (2) the number of morphosemantic relations after adding those specific to Bulgarian; and (3, 4) the number of morphosemantic relations learnt with the method described in Step 7 (Section 4.4). The results show that the extended method leads to an increase in coverage by 58.35% (compare the extended method (4) with the baseline (2)).

To assess the precision of the automatic relation assignment, we performed evaluation on five relations: *agent*, *undergoer*, *location*, *result*, and *state* (Table 3). The overall precision based on these relations is 0.774 for the baseline and 0.721 for the extended method, which shows that the performance of the method is relatively consistent.

We also obtained two types of generalizations based on WordNet and confirmed by the corpus

[8] `https://code.google.com/p/ws4j/`

[9] `http://lyle.smu.edu/~tspell/jaws/`

[10] `http://projects.csail.mit.edu/jwi/api/`

Relation	Precision (baseline)	Precision (extended method)
Agent	0.963	0.950
Undergoer	0.575	0.577
Location	0.857	0.750
Result	0.303	0.316
State	0.750	0.667

Table 3: Precision of the results for five semantic relations – baseline (Princeton and Bulgarian morphosemantic relations) and extended method (transfer of morphosemantic relations to pairs of nouns and verbs one of which does not participate in morphosemantic relations)

data that can be used for further classification. The first one represents the combinatorial properties of general semantic verb classes with particular (morpho)semantic relations. For example a verb of communication is more likely linked to an *agent* rather than to a *result* (Example 5).

Example 5. Frequency of relations in WordNet and the entire corpus.

verb.label	relation	fr wn	fr cor
verb.com	agent	744	555
verb.com	undergoer	306	362
verb.com	by-means-of	244	560
verb.com	result	192	283

Moreover, the nouns that are eligible to collocate as agents with a communication verb belong to a limited set of classes – `noun.person` or `noun.group` (Example 6).

Example 6. Frequency of relations in WordNet and the entire corpus.

verb.label	relation	noun label	fr wn	fr cor
verb.com	agent	noun.person	473	333
verb.com	agent	noun.group	271	220

The second generalization refers to the probability of the association of a given verb sense with a particular set of semantic relations and the general noun classes eligible for these relations. For instance, the communication verb *order* (Example 7) in the sense of *give instructions to or direct somebody to do something with authority* connects with the highest probability with an *undergoer* (`noun.person`) and an *agent* (`noun.person`).

Example 7. Relations of the verb *order* in WordNet and the entire corpus.

verb.label	relation	noun label	fr wn	fr cor
verb.com	undergoer	noun.person	33	8
verb.com	agent	noun.person	12	6
verb.com	by-means-of	noun.phen	9	7

6 Conclusions and Future Work

In this paper we have explored the applicability of the morphosemantic relations in WordNet for cross-language identification of semantic and in some cases syntactic dependencies between collocated verbs and nouns. As morphosemantic relations are valid cross-linguistically, the method is applicable for any language or a pair of languages.

The limitations of the proposed method lie in the insufficient connectivity of the nodes (synsets and literals). We have described an approach to automatic wordnet enhancement, which has resulted in a substantial increase in the number of morphosemantic relations. Another inherent weakness is that some of the relations are very general or vaguely defined. We have addressed this problem by considering relations jointly with the general semantic classes associated with the synsets in WordNet.

The method has the advantage of using limited linguistic annotation. It does not require text alignment, syntactic parsing or word-sense disambiguation. The cross-linguistic similarity partially disambiguates the target words, so that the senses for which the clauses are not similar are discarded; the semantic restrictions imposed by the general semantic classes and their compatibility also contribute to semantic disambiguation. The method is thus to a large extent language-independent and well suited to less-resourced languages.

In order to improve the performance and overcome the limitations of the method, we plan to explore deeper into the possibilities of predicting the roles of the verb participants from their general semantic class and the semantic compatibility of verb and noun classes, as well as from the compatibility of the different types of morphosemantic relations with the general semantic classes.

Another line of research to pursue in the future is the application of the proposed method and its subtasks to other NLP tasks, such as clause splitting, alignment based on wordnet relations, extraction of patterns from comparable corpora, and augmentation and enhancement of training data for MT.

References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. *Lecture Notes In Computer Science*, 2276:136–145.

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'12)*, pages 2596–2601.

Orhan Bilgin, Özlem Cetinouğlu, , and Kemal Oflazer. 2004. Morphosemantic Relations In and Across Wordnets – A Study Based on Turkish. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Global Wordnet Conference*, pages 60–66.

Xavier Carreras and Lluis Marquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2004*.

Xavier Carreras and Lluis Marquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*.

Christine Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting Semantics into WordNet's "Morphosemantic" Links. In Z. Vetulani and H. Uszkoreit, editors, *Proceedings of the Third Language and Technology Conference, Poznan, Poland. Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies*, volume 5603 of *Springer Lecture Notes in Informatics*, pages 350–358.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, September.

Jan Hajic. 1998. Building a syntactically annotated corpus: The prague dependency treebank. *Issues of Valency and Meaning*, pages 106–132.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antnia Martí, Lluís Márquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, page 305332. MIT Press.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, page 1933.

S. Koeva and A. Genov. 2011. Bulgarian language processing chain. In *Proceedings of Integration of multilingual resources and tools in Web applications. Workshop in conjunction with GSCL 2011, University of Hamburg*.

Svetla Koeva, Cvetana Krstev, and Dusko Vitas. 2008. Morpho-semantic relations in wordnet - a case study for two slavic langages. In *Proceedings of the Fourth Global WordNet Conference*, pages 239–254.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0(1):65–110.

Svetla Koeva. 2008. Derivational and morphosemantic relations in bulgarian wordnet. *Intelligent Information Systems*, XVI:359–369.

Joel Lang and Mirella Lapata. 2011a. Unsupervised Semantic Role Induction via Split-Merge Clustering. In *Proceedings of ACL 2011*, pages 1117–1126.

Joel Lang and Mirella Lapata. 2011b. Unsupervised Semantic Role Induction with Graph Partitioning. In *Proceedings of EMNLP 2011*, pages 1320–1331.

Claudia Leacock and Michael Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 265–283. MIT Press.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*.

Lluis Marquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006), Boston*.

Alessandro Moschitti and Cosmin Adrian Bejan. 2004. A semantic kernel for predicate argument classication. In *In Proceedings of CONLL 2004*, pages 17–24.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea*, pages 1399–1410.

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

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257.

Ted Pedersen, Serguei Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

Maciej Piasecki, Stanisaw Szpakowicz, and Bartosz Broda. 2009. A wordnet from the ground up. In *Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej*.

Maciej Piasecki, Radoslaw Ramocki, and Pawel Minda. 2012a. Corpus-based semantic filtering in discovering derivational relations. In *AIMSA*, pages 14–22.

Maciej Piasecki, Radosaw Ramocki, and Marek Maziarz. 2012b. Automated Generation of Derivative Relations in the Wordnet Expansion Perspective. In *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan, January.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Daniel Jurafsky. 2004. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of NAACL-HLT 2004*.

Sameer Pradhan, Wayne Ward, and James Martin. 2008. Towards Robust Semantic Role Labeling. *Computational Linguistics*, (34):289–310.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheczyk. 2010. Framenet ii: Extended theory and practice. Web Publication. `http://framenet.icsi.berkeley.edu`.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In A. Gelbukh, editor, *CICLing 2005, LNCS 3406*, page 100111.

Carina Silberer and Simone Paolo Ponzetto. 2010. UHD: Cross-lingual Word Sense Disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 134–137.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL-2003*, pages 8–15.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Márquez, and Joakim Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, pages 159–177.

Robert Swier and Suzanne Stevenson. 2005. Exploiting a Verb Lexicon in Automatic Semantic Role Labelling. In *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing 2005*, pages 883–890.

Kristina Toutanova, Aria Haghighi, and Christopher Manning. 2008. A global joint model for semantic role labeling. *Computational Linguistics*.

Piek Vossen. 2004. EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography*, 17(1):161–173, June.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP04, Barcelona, Spain*, July.