Towards the Automatic Identification of Light Verb Constructions in Bulgarian

Ivelina Stoyanova, Svetlozara Leseva and Maria Todorova

Department of Computational Linguistics, Institute for Bulgarian Language, BAS

CLIB 2016, 9 September 2016

Table of contents

1. Introduction
2. Properties of LVCs
3. Compilation of the training dataset
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Table of contents

1. Introduction
2. Properties of LVCs
3. Compilation of the training dataset
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Introduction

Objective

LVCs as a subclass of verbal MWEs with a view to their automatic recognition and annotation.

- An LVC consists of a light verb and a complement.
- LVCs' complement carries the predicative meaning of the MWE.

Introduction

Examples

- V+NP

  vzemam reshenie (*make a decision*)
- V+PP

  vlizam v kontakt (*come into contact*)
- V+A

  pravya lud (*make crazy*)
- V+Adv

  vzemam predvid (*take into consideration*)

Motivation

LVCs differ from free phrases:

- Their meaning is not fully decomposable to the sum of the meanings of their components.
- They are often not translated to other languages literally.

LVCs differ from idioms:

- LVCs have greater syntactic flexibility.
- LVCs are more semantically predictable.

Table of contents

1. Introduction
2. Properties of LVCs
3. Compilation of the training dataset
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Properties

LVCs are constructions with complex predication expressed by a verb and another predicative element.

- The light verb belongs to a small set of high frequency verbs with an abstract ('semantically bleached') meaning.
- The light verb expresses mainly aspect, directionality, aktionsart of the predicate.

Properties

- The complement is the semantic predicate and contributes the complex meaning of the expression.
- The complement denotes an abstract entity with eventive semantics.
- It is frequently expressed by a deverbal noun (*reshavam*$_V$ – *vzemam reshenie*$_N$).

Properties

- The nominal complement in many LVCs may vary in form – *vzemam reshenie/reshenieto (make a decision/the decision)*.
- It may take modifiers – *vzemam* **vazhno** *reshenie (make an* **important** *decision)*.
- The LVC allows for external elements, etc. – *vzemam* **barzo** *reshenieto (make* **quickly** *the decision)*.

Table of contents

1. Introduction
2. Properties of LVCs
3. **Compilation of the training dataset**
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Resources: the Corpus

- A subcorpus of the Bulgarian National Corpus containing news (10,655,068 words) and fiction texts (6,237,024 words);
- Linguistic annotation: sentence splitting, tokenisation, POS-tagging and lemmatisation using the Bulgarian LPC (Koeva and Genov, 2011).

Resources: BulNet

- We extracted a list of 2,239 verbal MWEs containing at least a verb and a noun;
- Their internal syntactic structure was determined as a sequence of POS tags – V NP or V PP (V P NP);
- The MWEs were manually categorised as LVCs or non-LVCs.

The MWEs so selected constitute the main part of the training data for the machine learning.

Resources: BulNet

- We extracted 105 verbs that can occur as part of LVCs fom BulNet – most are very ambiguous (15+ senses) and/or very frequent (in the Bulgarian National Corpus);
- Only 13 verbs have less than 5 senses and 4 verbs with less than 50 occurrences;
- None has both low frequency and a small number of senses.

Resources: BulNet

We extracted semantic information about LVCs' nominal components:

- Each verb and noun synset in WordNet has a semantic prime (unique beginners of the separate hierarchies in WordNet):
- We consider 10 of the noun primes, such as *noun.act, noun.state, noun.cognition* – those potentially expressing predicative meaning...

Resources: BulNet

- while excluding the remaining 15 noun primes, such as *noun.artefact, noun.person*.

The set of primes of all the possible senses of a given noun (with the application of some filtering procedures) were used as features in the machine learning.

Compilation of the training dataset

- The main part of the training dataset – 2,239 V NP and V PP MWEs classified as LVC or non-LVC
- 461 MWEs are identified as LVCs and the remaining as other types of verbal MWEs.
- To overcome the low number of LVCs and the lack of non-MWE instances, we extended the training set with additional data from the BulNC.

Compilation of the training and test dataset

- To this end we extracted verb-noun pairs with 10+ occurrences in the corpus. They were manually classified as LVC (true) and non-LVC (false) by two independent annotators. Only the instances in which the annotators agreed were taken into account.

Compilation of the training and test dataset

- We supplemented the training dataset so that it was increasd to 2,623 MWEs: 897 LVCs, the remaining – either non-MWEs or other categories of MWEs.
- The test dataset consisted of 200 unique candidates with 10+ occurrences extracted from the corpus in the same way as the additional training instances and annotated by the annotators into LVCs and non-LVCs, with equal number of both categories.

Table of contents

1. Introduction
2. Properties of LVCs
3. Compilation of the training dataset
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Lexical Features

- Certain LVs combine with certain nouns.
- The verb's lemma is defined as a feature.

Examples:

- *poemam (risk, otgovornost) – assume (risk, responsibility)*
- **vzemam (risk, otgovornost) – take (risk, responsibility)*

Semantic Features

- The semantic primes of the nouns extracted from BulNet.
- 10 (of the 25) noun primes relevant for predicative nouns: *noun.act, noun.cognition, noun.communication, noun.event, noun.feeling, noun.motive, noun.phenomenon, noun.process, noun.relation, noun.state*.

Semantic Features

- The labels for a noun are extracted and represented as a set, with additional procedures if the senses have different labels.
- If a noun's prime is not typical for predicative nouns, the respective sense is excluded from the nouns description.
- If in $>$ half of the senses a noun is non-predicative, it is ignored as a possible LVC component.

Semantic Features

Example: *vapros – (question, issue...)*: noun.act, noun.communication, noun.cognition, noun.attribute, noun.event

- *noun.act, noun.communication, noun.cognition, noun.event* (+)
- *noun.attribute* (–)

Statistical Features

- Information about the frequency of potential LVCs and their components in the corpus, i.e. the log-frequency of (a) the verb, (b) the noun, and (c) the LVC candidate.
- The association measure (MI) of the LVC candidate (used to determine between a collocation and an MWE).

Morphosyntacic Features

- The noun in many LVCs may take different forms (less typical for idioms): *vzemam reshenie* / *reshenieto* / *resheniya* / *resheniyata* (*make a decision* / *the decision* / *decisions* / *the decisions*).
- A binary feature is defined which takes **true** if the noun is found in more than one form in the corpus and **false** if the noun is invariable.

Syntactic Features

- Different word order: the feature takes the value **true** if 1+ variants are registered in the corpus, **false** otherwise.
- Components' modifiers: modifiers are limited to adjectives preceding the noun in the span of 2 tokens between the LV and its noun complement. The feature takes the value **true** if an example with a modifier is found in the corpus, **false** otherwise.

Syntactic Features

- External elements between the components: identified by their POS tag. The feature takes the value **true** when the POS tags of the elements (in the span of 2 tokens) are other than 'adjective' (possible modifier) or 'preposition' (PP head of a V PP LVC), **false** otherwise.

Derivational features

Predicative nouns tend to be of deverbal stems and hence derivationally related to a verb.

- The feature takes **true** if there is a V–N derivation, and **false** otherwise.

Examples: *nanasyam vreda (cause damage)*

the noun *vreda (damage)* is marked as derivationally related to the verb *vredya (to damage)* by matching the stem *vred-*.

Table of contents

1. Introduction
2. Properties of LVCs
3. Compilation of the training dataset
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Method outline

We trained and tested two classifiers on the feature set and the training set using two different learning algorithms based on decision trees – J48 and RandomTree (Hall et al., 2009).

Steps of LVC identification:

- Identify LVC candidates - the occurrences of a verb and a noun in the corpus with at most 2 tokens in between (except punctuation and conjunctions), taking into account the possibility for a free word order.

Method outline

- Filter the LVC candidates – remove candidates with low frequency in the corpus.
- Analyse the LVC candidates from the corpus in order to determine the variations in their form and word order, the possible modifiers and external elements separating the LVCs components.
- Apply the trained classifier to distinguish LVCs from other categories of phrases.

Table of contents

1. Introduction
2. Properties of LVCs
3. Compilation of the training dataset
4. Machine Learning Features
5. Method outline
6. Evaluation
7. Conclusions

Evaluation

We performed two-step evaluation:

- 10-fold cross-validation on the training set;
- evaluation on new test data of 200 unique LVC candidates extracted from the BulNC.

Evaluation

Algorithm	Precision	Recall	F_1
J48	0.739	0.794	0.766
RandomTree	0.710	0.741	0.725

10-fold cross-validation.

Evaluation

	Main method			Main method & Filtering		
Algorithm	Prec	Recall	F_1	Prec	Recall	F_1
J48	0.776	0.830	0.802	0.794	0.810	0.802
RandomTree	0.482	0.820	0.607	0.684	0.800	0.737

Evaluation on new test data.

Filtering: excluding candidates with association measure below the threshold of 2.0 (unlikely to be MWEs); excluding candidates with verbs that are not in the list of light verbs and/or nouns that do not belong to the predicative categories.

Table of contents

1 Introduction

2 Properties of LVCs

3 Compilation of the training dataset

4 Machine Learning Features

5 Method outline

6 Evaluation

7 Conclusions

Conclusions

- Performance comparable to methods based on similar features for other languages.
- The results emphasise the importance of semantic features such as the semantic primitive of the noun. The reduction of the noun primes to noun.act and noun.event shows that these are the most significant primitives and the precision result improves (0.782 with J48).

Conclusions

- The results reported in existing literature show that although LVCs seem to be a relatively well-defined class, their traits are not specific enough to distinguish them with higher precision from free phrases, collocations and idioms.
- Necessity to include more contextual and semantic features and to use the LVCs traits in a more productive way in engineering the ML features.

Thank you!

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

http://dcl.bas.bg/

Bulgarian National Corpus:
http://search.dcl.bas.bg/

Bulgarian Wordnet: http://dcl.bas.bg/bulnet/