Adding Linguistic Knowledge to NLP Tasks for Bulgarian: The Verb Paradigm Patterns

Ivaylo Radev
LMaKP
IICT-BAS
Sofia, Bulgaria
radev@bultreebank.org

Abstract

The paper reports pn a work on constructing automatically analytical paradigm of Bulgarian verbs on the bases of several existing language resources like Bulgarian inflection lexicon, Bulgarian Valency lexicons, BulTreeBank Bulgarian WordNet. The paper also discusses some possible usages of this new lexical resource containing Bulgarian verb paradigms and their English translations. This type of data can be used for machine translation, generation of pseudo corpora/language exercises, evaluation of parsers, and other tasks.

1 Introduction

The lack of training resources is a constant problem for many tasks within NLP. This is particularly true for languages like Bulgarian that are less resourced in some aspects. Automatically created labeled datasets are often seen as a solution to this problem. The creation of such data usually follows some kind of bootstrapping where the procedure starts with a set of seed elements and then an algorithm selects similar examples from a large corpus. Following this schema, the process could start with training a system on a small existing dataset and then analyze a large corpus from which new examples are to be selected; see for example (Mihalcea, 2002). Another strategy to produce automatically annotated data is to build pseudo corpora from existing resources; this is the approach applied for the creation of semantically annotated corpora from WordNets via Random Walk on Graphs algorithms (Goikoetxea et al., 2015). The algorithm for random walk on the knowledge graph of WordNet traverses the graph and emits a lemma and/or a word sense for each node respectively.

In our work, we produce syntactically correct sentences on the basis of several integrated resources for Bulgarian, including an inflectional lexicon, WordNet, a valency lexicon and a set of patterns for constructing the whole paradigms of Bulgarian verbs and the corresponding simple Subject + Verb + Indirect Object + Direct Object sentences exhausting all the possible word order alternations. In this paper, we demonstrate the patterns and the ways they can be used.

The Bulgarian verb is the grammatically richest part of speech (POS) of the language. The number of its synthetic forms goes up to 52. The analytical part of the verbal paradigm is much larger and comprises more than a thousand forms. Here we extend the paradigm to include not only verbal forms *per se* (simple forms, participles, auxiliary verbs and the particles да and ще) but also personal pronoun clitics for direct and indirect objects. Thus, for each verb we construct thousands of patterns which represent unique verbal forms. For example, a personal transitive verb like чета ("read") in present tense, 1st person, singular can be accompanied by one or, in this case, two clitics to form the following sentence:

(1) Чета им я .
Read-I them.DAT her .

Аз им я чета .
I them.DAT her read .

'I am reading it to them.'

One important characteristic of Bulgarian verbal forms is that they are in fact full-fledged simple sentences in their finite forms. Bulgarian is a pro-drop language and in most cases the direct and indirect objects can be optional as well. There are, of course, some exceptions to the rule. For example, the verb състоя се ("consist of") takes an obligatory indirect object. We rely on a valency lexicon of Bulgarian for presenting the selectional

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 76–82,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_012

restrictions of such cases. To sum up, we can generate thousands of simple sentences automatically, which in turn will benefit the creation of a lot of other resources. Even by itself, the dataset is valuable enough since it will contain patterns with up to 10253 verb paradigm members, including verbal complexes with all the possible combinations of subject, direct, and indirect object clitics, negative (няма, не), and interrogative particles (ли).

The paper is structured as follows: the next section discusses related work. Section 3 presents the extended verb paradigm (patterns sets). Section 4 surveys the possible impact of the paradigm data on NLP tasks. Section 5 concludes the paper and outlines future work.

2 Related Work

Previous efforts on adding linguistic knowledge to statistical machine translation for Bulgarian were done in (Simov et al., 2015). The paper reports on experiments done with machine translation from Bulgarian-to-English and English-to-Bulgarian under the project QTLeap.

The authors report problems with the so-called out-of-training word forms, word form pairs that do not appear in the parallel corpora used for the training. In order to solve this problem a parallel Bulgarian-English morphological lexicon was added to the parallel corpora. This lexicon was used in the POS tagging step, to provide all the possible tags for the known words and, in the lemmatization step, to convert each word form into its lemma.

The lexicon of 70 000 aligned word forms was constructed by exploiting several preexisting resources. First, word form lexicons for both languages were mapped to the corresponding part of the bilingual lexicon. Then, the corresponding word forms were aligned on the basis of morphological features like number, degree, definiteness, etc. This linguistic knowledge has been added gradually as factors in the MOSES system.

The paper reports a positive impact of the aligned word form parallel lexicon on the translation in both directions, but the addition of the definite forms for English did not change the result.

The lexical resource WordNet (WN) has established itself as one of the most used and popular language data resources in the field of NLP. WN can be described as a kind of thesaurus that groups word meanings or senses together and labels the semantic relations among them.

The BulTreeBank Bulgarian WordNet (BTB-WN) — (Simov et al., 2019) is a newly created and expanding lexical resource for Bulgarian language. It currently contains 22 000 synsets manually mapped to the Princeton WordNet (PWN) and continues to grow due to the process of linking it with the Bulgarian Wikipedia. The role of the BTB-WN is to provide lexical and semantic data for NLP tasks for Bulgarian such as sense disambiguation (WSD), relation extraction, named entity and multiword expression (MWE) parsing, machine translation, etc.

One example for experimentation with WN is (Mihalcea, 2002). The paper describes an algorithm for the automatic generation of GenCor, a large sense tagged corpora, for participation in SENSEVAL-2. The generation algorithm works in three steps: (1) creation of a set of seeds (sense tagged examples from SemCor; WN and rule creation); (2) searching in the Web with the seed expressions; (3) disambiguation of words in a small text snippet surrounding the seed expressions.

The idea of creating sense tagged examples out of WordNet is based on the assumption that each example and its corresponding synset are properly linked, which allows to assign the correct sense to at least one word in the examples. The relations between words taken into consideration are identity, synonymy, hypernymy, hyponymy, and sibling terms.

The usage of WN in recent years and efforts to link it with other resources (BabelNet; UBY) show that it is beneficial to use multiple language resources at once, especially for low-resource languages that do not have such resources or their existing resources are small in size.

Another grammatical data resource used in NLP tasks for Bulgarian is the valence lexicon presented as part of the Bulgarian Ontology-based Lexicon (Osenova et al., 2012). The lexicon exploits the relation between ontology and text. This lexicon is mapped to an ontology in order to connect lexical units to their conceptual meanings. Additionally, the lexicon contains phonological, morphological, and syntactic linguistic knowledge.

A related paper (Osenova and Simov, 2015) reports that the lexicon contains 4113 valency frames coupled with the respective meanings and that it covers 1903 lemmas. It considers the verbs as the most important part of speech for the task of semantic role annotation.

The valency frames are extracted from the BulTreeBank, manually linked with verb senses and detailed participants with respect to the usage, and then returned back into the treebank (Osenova et al., 2012). This ensures that the sense and the frame are appropriate for the respective usage for each verb occurrence in the treebank. The semantic classes of the verbs are transferred by the mappings of the Bulgarian valency lexicon to the PWN, which, together with the valency frames, helps in the process of selection of the appropriate semantic roles. After that the semantic roles are transferred to the corresponding constituents in the tree of the verb occurrence.

3 The Verbal Paradigm Patterns

In this section we present the types of patterns which are used for the generation of all members of the extended verbal paradigm. In order to generate all of these forms we create patterns that include the verb synthetic form, clitics, auxiliaries, etc.

From all the parts of speech in Bulgarian, the verb bears the most information. It contains grammatical information not only about the predicate expressing an event, but also for the participants in this event. The grammatical characteristics of the verb are: 9 tenses (1 present tense, 4 past tenses and 4 future tenses); perfective and imperfective aspect; singular and plural number; first, second and third person; gender in the participle forms; active and passive voice (although some argue for one more — reflexive); indicative, imperative, conditional mood, and three evidentials: renarrative, dubitative, and conclusive. Thus the patterns represent the allowed combinations of these forms and features. Each pattern for a given form consists of a form of the main verb and some auxiliary elements which include auxiliary verbs as well as some verbal particles. Because we want to express also the negative, interrogative, and passive voice forms, we include such forms in the verbal paradigm patterns. The last element of the extended verbal paradigm is the valency potential of the verb. Here we assume only the internal arguments of the verb — the subject, the direct object, and the indirect object. All of them could be represented via nominative, accusative or dative clitics. These clitics can be in singular or plural number, and in first, second, or third person. Additionally, we create a pattern for each possible word order of the corresponding main verb form, auxiliary, verbal particles and clitics. In our work we consider a verb form to be determined by its grammatical characteristics. Its realization based on omitted pronouns (clitics) or movement in the word ordering of the particles and pronouns is called variation.

All this results in many forms and variations. The extended paradigm of the verb чета ("read") contains 1205 verb forms and 10253 variations with explicit subject, direct, and indirect object clitics.

The initial idea behind the construction of a Bulgarian verb paradigm pattern set was for it to be used in the improvement of the coverage of the Bulgarian treebank. The motivation for this is that only a small percentage of the verb forms could be found in the available corpora of Bulgarian. For example, the form Някой чете нещо "Someone reads something" is basically omnipresent and the form Щял съм бил да им я чета "(they doubt) I would be reading it to them" is very rarely attested in everyday (web) language. We have created verb paradigm pattern sets for nine types of Bulgarian verbs — see Table 1. These types of verbs are described by the grammatical features of their stems and the number of the paradigm members vary for each type of verb.

The representative verbs for each type were selected randomly to cover basic grammatical information for: personal/impersonal verbs; transitive/intransitive verbs; reflexive verbs and the perfect/progressive aspect of verbs. All of the paradigm patterns are encoded manually for the representative verb of the corresponding type. Additionally, each lexical item in each pattern receives its POS tag from the BulTreeBank tagset — (Simov et al., 2004). Also, the lexical items in the patterns are trivially lemmatized.

As it was mentioned, each of these lemmas is conjugated in all possible verb forms for tense, person, number, mood and voice. The clitics for subject, direct and indirect object are added. The forms also include tree more variations: negation, question and a combination of the two. In some cases more than one word ordering are possible. The negation variants are formed with the particle не "not". For example:

No	Verb	Features	Transcription	Translation
1	може	impers; intr; ipfv; mod	'mozhe'	can
2	трябва	impers; intr; ipfv; mod	'tryabva'	have to
3	вървя	pers; intr; ipfv	'varvya'	I walk
4	чета	pers; tr; ipfv	'cheta'	I read
5	прочета	pers; tr; pfv	'procheta'	I read (it all)
6	сърби ме	pers; intr; ipfv; acc	'sarbi me'	It is iching me
7	домъчнява ми	impers; intr; ipfv; dat	'domachnyava mi'	(I) start to feel grief (for something)
8	смея се	pers; intr; ipfv; refl	'smeya se'	I am laughing
9	изсмея се	pers; intr; ipfv; refl	'izsmeya se'	I am laughing (once)

Table 1: The current verbs in the paradigm resource. Grammatical features: impers = impersonal, pers = personal, tr = transitive, intr = intransitive, pfv = perfective, imperfective = ipfv, refl = reflexive, mod = modal, dat = dative clitic verb, acc = accusative clitic verb.

(2) Не им я чета .
Not them.DAT her read-I .
'I am not reading it to them.'

The interrogative variants are formed with the interrogative particle ли. For example:

(3) Чета ли им я ?
Read-I INTER them.DAT her ?
'Am I reading it to them?'

Finally, the combination of negative and interrogative variants has some possible word orders:

(4) Не им я чета ли ?
Not them.DAT her read-I INTER ?

Не им ли я чета ?
Not them.DAT INTER her read-I ?

'Am I not reading it to them?'

As was mentioned above, each of the variations is also a plausible simple sentence in Bulgarian. There are a few exceptions like participles that can be used only in attributive constructions and gerunds.

Finally, possible translations to English are included after every form. It is important to note that the two Bulgarian aspects are considered different lemmas and the Bulgarian language does not use continuous tenses as the English does. A present continuous tense does not exist in Bulgarian. Both languages have imperfect tenses, but only in name. In Bulgarian, the perfective and imperfective aspects have forms for imperfect tense. These dissimilarities lead to variations in the translations of the tenses. The translation patterns depend only on the forms of the main verb and its translation into English.

Up to here we have presented the construction of verb paradigm pattern sets for the nine main types of Bulgarian verbs. In order to apply them to arbitrary verbs we need to link the patterns with other language resources. More precisely to an inflectional lexicon, a valency lexicon and a Bulgarian WordNet. Each of these resources provides pieces of the puzzle that are necessary for the application of the patterns.

The first step is to determine the paradigm types via mapping the paradigm pattern type to the verb type. Data for the verb types will come from the inflectional morphological lexicon. On the basis of the grammatical features we select the correct verb pattern set. For example the lemma of the verb дарявам ("to gift") has the same POS tag as the verb чета ("read") and is also transitive and imperfective. Thus from the inflectional lexicon we receive the synthetic paradigm of the verb and its grammatical features of the stem.

The next step is to extract information about the possible clitics of the verb. This information is available within the Bulgarian valency lexicon. From the pattern set and the POS tags we know that чета ("read") has dative and accusative clitic. Then we need to check the frame for дарявам ("to gift") if it can also have direct and indirect object to transform the pattern:

(5) Чета ли им я ?
Read-I INTER them.DAT her ?

(6) Дарявам ли им я ?
Gift-I INTER them.DAT her ?

The last necessary bit of information is the English translation, which we find within the BTB-WN. As was presented above, the valency lexicon was integrated with BTB-WN. Thus, when we select a Bulgarian verb together with its inflectional type and valency frame we also determine its potential senses within BTB-WN. The mapping from BTB-WN to the English WordNet is used to select the English verb.

Utilizing all this information, we could construct the whole extended paradigm of the selected verb and the corresponding translations in English:

(7) Чета ли им я ?
Read-I INTER them.DAT her ?
'Am I reading it to them?'

(8) Дарявам ли им я ?
Gift-I INTER them.DAT her ?
'Am I gifting it to them?'

4 Application of the Extended Verbal Paradigm

In this section we present some applications of the generated extended verbal paradigms. Some of these applications require extensions of the patterns in order to add the necessary linguistic knowledge to the verbal forms.

The immediate NLP applications of the new language resource include POS tagging and lemmatization. Although we have the rules by which the verbal forms are generated and we could easily turn them into an analytical module, the resource could be used for training and testing statistical or neural network POS taggers. Because most of the clitics and many of the verbs are ambiguous, the task of POS tagging is not trivial.

Another obvious application is in the area of statistical and neural network machine translation, similarly to the experiments reported in (Simov et al., 2015). We hope that in this way the MT system would be able to learn to translate analytical verbal forms.

In order to support other NLP tasks we need to extend the resource with more linguistic knowledge. To support dependency parsing we need to convert each verbal form which represents a sentence into Universal Dependency (UD) format. This is straightforwardly done via rules for each of auxiliaries, clitics and particles. For example the sentence from above:

(9) Аз им я чета .
I them.DAT her read .
'I am reading it to them.'

is converted to the following UD tree depicted in Figure 1 for the example 9.

After converting the extended paradigms into UD trees provides an useful resource for training and testing UD parsers. But it is obvious that the utilities of simple sentences comprising a verb, auxiliaries, particles and clitics is not huge. In order to make them really useful we need to include also full-fledged arguments. In order to do this we need to extended the patterns with positions of the full-fledged arguments with respect to the other components of the verbal forms.

Then using the mapping from the main verb to the valency lexicon we could determine the sense annotation of the arguments of the verb. These senses are linked to appropriate synsets in BTB-WN. This allows to select appropriate lemmas for each argument. Then having grammatical features for each argument stated in the verbal form we could generate the correct word form for the arguments.

If for the verb чета we have the notion that an "agent" can read an "information object", we can substitute the pronouns with full words. In this way we convert the sentence:

(10) Той им я чете .
He them.DAT her read .
'He is reading it to them'

into the sentence:

(11) Учителят чете книга на учениците .
Teacher-the read book to students-the .
'The teacher is reading book to the students.'

Another kind of data that can be imported from WN comes from its "instance-of" relation for generating sentences with named entities:

(12) Барак Обама чете Властелинът на
Barack Obama read lord-the of
пръстените на учениците .
rings-the to students-the .
'Barack Obama is reading the Lord of the rings to the students.'

We can also use the mapping of BTB-WN to PWN to translate the positions in the pseudo sentences bidirectionally from Bulgarian to English and from English to Bulgarian. This will be an even better source of parallel data for machine translation models.

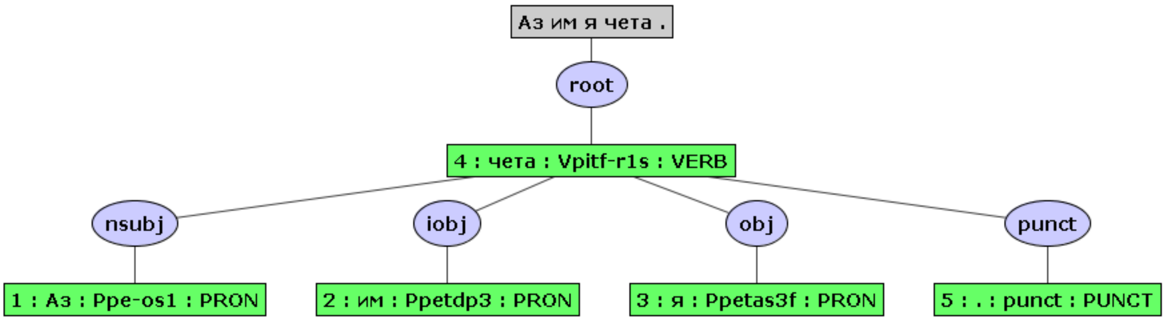

Figure 1: UD tree for the example 9.

It is also easy to extend the conversion module in order to represent such sentences into UD format. The other consequence of the procedure for the generation of sentences is that we know the senses of each word in them. In this way the new sentences could be used for training and testing UD parsers and modules for the Word Sense Disambiguation task.

The Bulgarian language (as a Balkan one) also uses clitic doubling:

(13) Учителят им я чете книгата на
Teacher-the them.DAT her read book-the to
учениците .
students-the .

'The teacher is reading a book to the students.'

This phenomenon is rarely seen in corpora, but it is used in everyday communication. It allows for logical emphasis and relates contrast. In cases where the head of the object is in front of the predicate the doubling is mandatory. It may be used for both direct and indirect objects.

The good thing about this kind of transformations is that the information for each position in the string is known in advance and everything generated by the automatic system using these resources will have a morpho-tag, lemma, UD annotation, sense disambiguation and translation to English. Another benefit is the control over the parameters for generation; the process can be tuned to get data for specific task.

Corpora containing data from news media, web crawlers and social networks often do not cover all of the linguistic knowledge for a given language. We need pseudo corpora that add this missing information for the training and evaluation of natural language parsers. This kind of pseudo corpora can be generated automatically. The automated method for the generation of training and evaluation data is a core one in the field of NLP and it has been in use for many years.

A resource consisting of sentence strings that combine morphological information, verb frames and sense annotation can be used as the basis for rule-based generation of Universal Dependencies trees. The combination of word sense and verb will provide data for restricting the agents and the positions for direct and indirect objects. This can be done first for Bulgarian and later for English.

5 Conclusion and Future Work

In this paper we present the construction of extended verbal paradigms. The integration of these resources with other language resources like a valency lexicon, BTB-WN and a morphological lexicon converts these paradigms into a well annotated corpus of simple sentences. Thus, the verb paradigm patterns show promise for positive impact on various NLP tasks. The future work on linking it to other linguistic data resources will allow for more specific experiments to be conducted.

One criticism of the approach for constructing full-fledged sentences is that the selected full-fledged subjects, direct and indirect objects are retrieved from WordNet quite randomly. In this way the resulting sentences are far from natural ones. In order to address this problem, in the future we envisage to extract examples of co-occurrences of subjects and objects from automatically parsed corpora and to experiment with the extracted phrases to generate new sentences.

Another task is to experiment with reverse parsing. For instance, taking one sample sentence from real text corpora and transforming it into a new sentence with a rarer verb form. We expect to be able to convert sentences like this one:

(14) Учителят им чете книга .
Teacher-the them.DAT read book .
'The teacher is reading a book to them.'

to sentences like this one:

(15) Бил ли им е чел
was INTER them.DAT is read.PTCP.SG.M
учителят книга ?
teacher-the book ?
'Has the teacher read a book to them?'

Our next task will be to evaluate experimentally the usefulness of this new resource. We plan to perform experiments for each of the tasks: POS tagging, UD parsing and WSD.

Acknowledgments

This work was partially supported by the *Bulgarian Ministry of Education and Science under the National Research Programme "Young scientists and postdoctoral students" approved by DCM # 577 / 17.08.2018* and by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-164/28.08.2018.

References

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *HLT-NAACL*, pages 1434–1439. The Association for Computational Linguistics.

Rada F. Mihalcea. 2002. Bootstrapping large sense tagged corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Petya Osenova and Kiril Simov. 2015. Semantic role annotation in bultreebank. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 148–156, Warsaw, Poland. TLT14 2015.

Petya Osenova, Kiril Simov, Laska Laskova, and Stanislava Kancheva. 2012. A treebank-driven creation of an ontovalence verb lexicon for Bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2636–2640, Istanbul, Turkey. LREC 2012.

Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, and Zara Kancheva. 2019. Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. In *Proceedings of the 10th Global WordNet Conference*.

Kiril Simov, Petya Osenova, and Milena Slavcheva. 2004. BTB-TR03: BulTreeBank Morphosyntactic Tagset. Technical report, Bulgarian Academy of Sciences.

Kiril Simov, Iliana Simova, Velislava Todorova, and Petya Osenova. 2015. Factored models for deep machine translation. In *Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015)*, pages 97–105.