An Event-centered Knowledge Graph for Bulgaria

Kiril Simov[1 [0000−0003−3555−0179]], Nikolay Paev[2,3 [0009−0006−6125−2684]], and Petya Osenova[3 [0000−0002−4484−5027]]

Institute for Information and Communication Technologies
Bulgarian Academy of Sciences
https://www.iict.bas.bg/

Abstract. The paper presents the principles behind the event-centered knowledge graph for Bulgarian. The methodology and technology for the construction of the knowledge graph are outlined together with the integration and linking of the relevant knowledge-bearing resources. In addition, some first experiments with a trained T5 large language model (LLM) for detecting events and their roles have been performed. The best results correlate with the most frequent events. Some preliminary qualitative evaluation has been conducted as well.

Keywords: event corpus · knowledge graph · LLMs · Bulgarian

1 Introduction

In order to support research within social sciences and humanities (SS&H), we need to provide information management about a huge variety of research objects, including different kinds of texts (various genres, domains, time periods), artefact models, art masterpiece representations and descriptions, building construction documentation, pictures, etc. The usual unification of these data is the metadata, but the common information represented in this way turns out to be insufficient. At the other end of the scale there exists very specific data and tools for their management such as creation (digitization), representation, generalization, search, etc. We consider identification of information of interest and its simultaneous observation within the same context to be one of the steps when conducting research within the SS&H. In order to support this type of research in SS&H, we would require to put the varying types of data in the context of each other. This requirement is called a *contextualization*.

The main characteristics of the contextualization are *time* and *space*. Thus, it can be observed which events happened at the same time or in the same space. The additional characteristics refer to the participants in the specific events; the similar constitution or appearance, etc. of physical objects like form, size, material; the similar style of representation in images, texts and sounds; the same school of production, etc.

When metadata is excluded (which is obviously part of the context description), most of the information within the individual datasets of SS&H resources does not contain enough contextual information for appropriate integration with

other existing knowledge datasets. For that reason, we need to construct a new layer of information necessary to support the contextualization. As a starting point, we focused on the following main kinds of entities: **People** — their biographies – their characteristics, motivations, opinions, events in their lives, roles they played; **Objects** —- geographical ones, artefacts, etc. and their features; **Events** —- event type, place, time, participants (People, Events, Documents, Objects, etc.), relations to other events; **Documents** – authors, content, opinions, events, entities; etc.

We consider text as the main source of information for the represented objects. From a methodological and technological points of view, we consider the usage of Linked Data and the representation of the information as a knowledge graph — [1]. Since the main body of knowledge represented in it is related to Bulgaria, it is called a *Bulgaria-centric Knowledge Graph* (BGKG).

2 The Main Components of BGKG

We consider BGKG to consist of several main components: a Bulgarian Wordnet (BTB-WN), a Valency Lexicon of Bulgarian (BTB-VAL) and a BGKG(I) — a section of BGKG which consists of the instances. We consider BGKG as an event-centric knowledge graph that follows [6]. They define events as *things that happen*, comprising four components: (1) an event action component describing what happens or holds true; (2) an event time slot anchoring an action in time that describes when something happens or holds true; (3) an event location component specifying where something happens or holds true; and (4) a participant component that gives the answer to the question: who or what is involved with, undergoes change as a result of or facilitates an event or a state. Thus, the relations between objects are described in terms of events in which they participate. The BTB-Wordnet and BTB-Valence Dictionary are part of a Bulgarian Language Integrated Language Resources Tool Kit. They represent the language component of the ontology constructed for BGKG — an extension of the CIDOC-CRM ontology[1]. We selected this ontology because it is related (at least partially) to our domain of interest. It supports the appropriate formalization of the events, activities and states. This top ontology of events allows for an easy extension with new event classes following a frame-semantic approach. In addition, CIDOC-CRM has many extensions for different domains related to SS&H. In our work the properties of each event class in CIDOC-CRM are called *roles of the participants in the event.*

2.1 The Bulgarian Wordnet and Valency Lexicon

BTB-WN ([9]) is a WordNet of Bulgarian, constructed along the lines of the Open English Wordnet[2] — [3]. The words are represented in synonymic sets

[1] https://cidoc-crm.org/
[2] https://en-word.net/

(called synsets). Each synset represents one meaning per lemma in it. The meaning is described by a gloss and it is illustrated by example usages. It can be assumed that each synset represents a concept. The set of synsets is organized within a network through semantic and lexical relations like hyperonymy, hyponymy, meronymy, antonymy, etc. BTB-WN is aligned with the Open English Wordnet by two semantic relations `equivalent-to` and `near-equivalent-to`. The first one is used when the Bulgarian and English synsets are completely equivalent to each other. The second is used when the Bulgarian synset is equivalent to a union of several English synsets. In addition, the hyperonymy, hyponymy, and similarity relations are used. We also support a mapping between the BTB-WN synsets and the articles within the Bulgarian Wikipedia. The mapping has been performed for more than 11 000 synsets. The vocabulary in BTB-WN has been selected to cover a number of important existing vocabularies: from English: Core WordNet[3] (5 000 more frequently used word senses); Global Wordnet Association's Base Concepts[4] (about 5 000 synsets); from Bulgarian: a Semantic Minimum Dictionary of Bulgarian (1 712 senses); a Bulgarian-English Dictionary (7 893 lemmas); a List of lemmas within the Bulgarian treebank – BulTreeBank (BTB) (near 9 000 lemmas); Vocabularies of several textbooks for studying Bulgarian at levels A1-A2; B1-B2, C1 (near 5 000 lemmas), etc. The selection of the new words follows the BTB Frequency List. In this way, the coverage of BTB-WN was extended. Currently BTB-WN has more than 36 000 synsets. Thus, the most important Bulgarian senses are included in the resource.

BTB-VAL is a Bulgarian valency lexicon constructed on the basis of annotations from the BulTreeBank. The verbs with their arguments were extracted from the treebank and classified by their senses within BTB-WN. Then the arguments were also mapped to the corresponding synsets. Thus, one syntactic frame could result in several semantic frames. The opposite is also true — each description of an event in CIDOC-CRM (or its extension) might be related to more than one valency frame in BTB-VAL (currently, 4 555 frames).

2.2 Bulgaria-centric Knowledge Graph — Instances

The construction of the BGKG(I) is being performed in two major steps: analysis of the infoboxes (InfBs) of Bulgarian Wikipedia similar to other knowledge graphs like YAGO ([10]) and DBpedia ([2]), and analyses of appropriate domain texts. The extraction of the information from the infoboxes is performed semi-automatically from the XML dumps and the HTML formatting of the articles in the Bulgarian Wikipedia. First, we extract all the words from InfBs and extend BTB-WN to include appropriate senses for them. In this way, it is guaranteed that the correct meaning is used for the access to the conceptual knowledge in the ontology and the instance data. For each type of object (like persons, countries, settlements, geographical objects, organizations, etc.), the characteristics were examined that are mentioned in their InfBs and appropriate events

[3] https://wordnet.princeton.edu/download/standoff-files
[4] https://globalwordnet.org/resources/gwa-base-concepts

were defined to be added within the ontology. The definitions of these events are created through the mapping to the related valency frame in BTB-VAL. This ensures also a mapping to BTB-WN and then to the events in the ontology. The size of BTB-VAL is large enough to support the mapping between the linguistic definition of the event and its valency frame. At the same time, the mapping to the ontology is more problematic. It might require an extension of the ontology in the part of the events and in some cases also for the object classes. The mapping to the ontology starts with the most specific events in the existing version of the ontology. If they are appropriate for the current version of the knowledge graph we only record that the events need to be made more specific in future.

From each article with an infobox a description of an instance is constructed. The URL of the instance is the URL of the Wikipedia page, which will be made specific to BGKG. For each row in the infobox we manually assign events already extracted at the previous step of the processing. The rows in the infobox contain information about the roles of the corresponding events. Thus, from an InfB we extract the core facts about the corresponding entities.

The next steps in the process of constructing the BGKB is to extract appropriate events from domain texts. For this task we had to annotate an event corpus of Bulgarian and to prepare the necessary language pipeline for Bulgarian. In the next section we briefly present the corpus, the existing pre-processing pipeline and the LLM-based experiments for the task of the event extraction.

3 Language Technologies for BGKG

Here we present the existing language resources and the trained models for text analysis with the aim to extract event knowledge.

3.1 Linguistic Pre-processing of Bulgarian Texts

The above mentioned *Bulgarian Language Integrated Language Resources Tool Kit* consists of the central lexicons and corpora necessary for the processing of Bulgarian. It includes an Inflectional dictionary, a morpho-syntactic corpus, a treebank, several Named Entities corpora with annotated and linked to Wikipedia names, BTB-WN, BTB-VAL and some other resources. These are integrated at different levels – through the lexical, sense and syntactic structures. Where it was possible, the annotations were performed over the same texts. We have pretrained a number of in-house LLMs to solve a wide range of basic Natural Language Processing (NLP) processing taks: Tokenization, Part-of-speech tagging, Lemmatization, Dependency parsing — [7], Word Sense Disambiguation — [8], Named Entity Recognition and Linking. These serve as a good starting point for the event extraction task. In order to implement such a model, we exploit the Bulgarian Event Corpus — [4]. The annotation schema is based on CIDOC-CRM (with extensions) where the events are modeled as a tuple of a type and roles. For example, the event of birth is represented as: ⟨ Type : **Birth**, Roles : ⟨ **brought-into-life** (the new born person), **mother**, **father**, **place** (the

birth place - usually the name of a city, country or hospital), **time** (the time of birth - usually it's a date, but can include hours or it's just month and year) ⟩ ⟩. The values of the roles are assumed to be NEs. Thus, the event annotations facilitate the extraction of instance information from texts. In the rest of this section an LLM-based model is discussed for the event extraction task.

3.2 An LLM for Event Extraction

We model the annotation of events in a sentence as a *text-to-text* task. The input is the sentence, and the output is a list of events, each of which has a type and a set of roles. The roles also have types and spans of text from the original sentence. The output is formatted as JSON files to make it easier to parse. The next paragraphs describe the model pretraining and fine-tuning for the event extraction task.

Training. As a base model to fine-tune, we consider our own pre-trained T5 model with 403M parameters. The pre-training has been done for 3 epochs on a corpus of 20B Bulgarian tokens. The training objective was span denoising (as introduced in [5]) with 0.25 noise density and a mean noise span length of 3 tokens. We fine-tuned for 10 epochs on a train split of the event corpus sentences. The best checkpoint was selected based on the loss over a validation set. The model learned well to output correct JSONs – the invalid outputs on the test set were only 0.2%.

Quantitative analysis of the results on the test set. The quantitative analysis of the results over the test set is not trivial. Let us consider two lists – of reference events and predicted events. We can perform some shallow comparison measuring only the overlap between the types of the events, or deep comparison trying to ensure pairing between the reference and the predicted events, and then make a shallow or deep comparison over the roles.

To keep it simple, we only calculated the precision and recall metrics over the types of the events as well as the precision and recall metrics over the types of the roles after greedy pairing of predicted and reference event types.[5] The results are presented in Table 1. The precision of the event types measures the number of the predicted events/roles that are also present in the reference events. On the other hand, recall denotes the number of the reference events/roles that are present in the predicted events. Similarly to the greedy pairing of events, one can do a greedy pairing of the roles in an event, and can measure the BLEU metric between the predicted text spans. The result is 71.67%.

The average count of reference events from the sentences in the test set is 0.7070, while the average count of predicted events is 0.5295. That aligns with the higher precision of the model compared to the recall. Table 2 gives a summary of the recall for some of the events in the event corpus.

[5] The first predicted event with type A is paired with the first reference event with type A, and then analyses are made on the types of the roles. Please note that this step is not correct all the time, especially if there are two events with the same type in a single sentence.

Table 1. Precision and recall of the types of the predicted events and the types of the event roles after greedy matching between the reference and the predicted events

	Precision	**Recall**
Predicted event types	75.14%	56.28%
Predicted role types of the correct events	89.22%	87.20%

Table 2. List of the best recalled event types.

Event	**Results**
Purchase-sale	1.0
Birth	0.9375
Education	0.8776
Kinship	0.8718
Death	0.8235
Founding	0.7778
Relocation	0.6939

Qualitative analysis of the results on the test set. There are four types of errors. These are: a) an unrecognized event with regard to the training data, b) an unrecognized role with regard to the training data, c) a wrongly recognized event with regard to the training data and d) a recognized event or a role that is missing in the training data.

With regard to a) might be due to the existence of sometimes longer triggers which induces the scarcity of the necessary examples for good training. Also, when the trigger is a verb form, its occurrence in the data might drop drastically. With regard to b) the following cases are identified: i) unidentified adverbials (Time, Manner, Purpose, etc.), expressed by prepositional phrases, and ii) too specific roles such as *Area of somebody's activity*. As for c), the main problem is that only one event has been used as a reference. Thus, in the multi-event sentences the model can predict just one from all the events that were not used as references. Thus, we leave the multi-input scenario with LLMs for future work. Concerning d), it is often observed that the role has not been identified as the correct one since the description related to it was partial.

4 Conclusion

In this paper we presented our approach for constructing a linguistically-aware BGKG. We first extracted instance information from the infoboxes of Wikipedia. This information was aligned with the BTB-Wordnet and the Valency lexicon. This step plays a double role – it would facilitate the interaction between the users and the BGKG, but it also would support the extraction of new information from texts. We also presented a model for event extraction. In future, we plan to work in the direction of extending the inventory of events in the ontology and improving the linguistic processing pipe for better extraction of conceptual data.

Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology – https://www.cost.eu. The reported work has been supported by CLaDA-BG, *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.*

References

1. Hogan, A., Blomqvist, E., Cochez, M., D'amato, Cl., Melo, G. De, Gutierrez, Cl., Kirrane, S., Gayo, José E. L., Navigli, R., Neumaier, S., Ngomo, Axel-Cyrille Ng., Polleres, Ax., Rashid, S. M., Rula, An., Schmelzeisen, L., Sequeda, J., Staab, St., and Zimmermann, An.: Knowledge Graphs. ACM Computing Surveys **54**(4), 1–37 (2021).
2. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Kleef, P.V., Auer, S., and Bizer, C.: DBpedia — A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal **6**(2), 167–195, (2015).
3. McCrae, J. P. , Rademaker, Al., Bond, Fr., Rudnicka, E., and Fellbaum, Ch.: English WordNet 2019 – An Open-Source WordNet for English. In Proceedings of the 10th Global Wordnet Conference, 245–252, (2019).
4. Osenova, P., Simov, K., Marinova, I. and Berbatova. M.: The Bulgarian Event Corpus: Overview and Initial NER Experiments. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 3491–3499, (2022).
5. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, Sh., Matena, M., Zhou, Y., Li, W., and Liu, P. J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, **21**(1), Article 140, 67 pp, (2020).
6. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T.: Building event-centric knowledge graphs from news. Journal of Web Semantics **37–38**, 132–151, (2016).
7. Paev, N., Simov, K., and Osenova, P.: Introducing Shallow Syntactic Information within the Graph-based Dependency Parsing. In Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024), pages 46–54, (2024)
8. Paev, N., Simov, K., and Osenova, P.: Word Sense Disambiguation with Large Language Models: Casing Bulgarian. 13th International Global Wordnet Conference (GWC 2025), (2025).
9. Simov, K., and Osenova, P.: Recent Developments in BTB-WordNet. In Proceedings of the 12th Global Wordnet Conference, pages 220–227, (2023).
10. Suchanek, F. M., Kasneci, G., and Weikum G..: YAGO: A Large Ontology from Wikipedia and WordNet. Journal of Web Semantics **6**(3), 203-–217, (2008).