Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3491–3499
Marseille, 20-25 June 2022

The Bulgarian Event Corpus: Overview and Initial NER Experiments

Petya Osenova, Kiril Simov, Iva Marinova, Melania Berbatova
Artificial Intelligence and Language Technology, Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Bulgaria
petya@bultreebank.org, kivs@bultreebank.org, iva.marinova@identrics.net, melania.berbatova@gmail.com

Abstract

The paper describes the Bulgarian Event Corpus (BEC). The annotation scheme is based on CIDOC-CRM ontology and on the English Framenet, adjusted for our task. It includes two main layers: named entities and events with their roles. The corpus is multi-domain and mainly oriented towards Social Sciences and Humanities (SSH). It will be used for: extracting knowledge and making it available through the Bulgaria-centric Knowledge Graph; further developing an annotation scheme that handles multiple domains in SSH; training automatic modules for the most important knowledge-based tasks, such as domain-specific and nested NER, NEL, event detection and profiling. Initial experiments were conducted on standard NER task due to complexity of the dataset and the rich NE annotation scheme. The results are promising with respect to some labels and give insights on handling better other ones. These experiments serve also as error detection modules that would help us in scheme re-design. They are a basis for further and more complex tasks, such as nested NER, NEL and event detection.

Keywords: Bulgarian event corpus, multidomain dataset, NER experiments, Social Sciences and Humanities

1. Introduction

In recent years there is increasing interest in event-oriented applications within the NLP community for variety of tasks using different methods (like supervised or unsupervised ones). Such applications are event detection and extraction, scheme induction, event profiling, among others. The successful extraction of events is not trivial but it became possible thanks to the advancements in technology (improved language modeling as well as cross-lingual and cross-domain methods), on the one hand, and also thanks to the availability of a lot of hand-made, hand-induced or automatically produced data, on the other.

In this paper we present our work on the creation of a Bulgarian Event Corpus (BEC) and the challenges behind that. We also report the results from the initial experiments with NER models trained on it.

The main reason for the construction of the BEC corpus is to have appropriate data for training Named Entity Recognition (NER), Named Entity Linking (NEL) and Event Recognition models. Such models would help us in the extraction of structured knowledge from domain texts in the area of Social Sciences and Humanities (SSH). The extracted structured knowledge will be ultimately used for the creation of a Bulgarian-centric Knowledge Graph — see (Simov and Osenova, 2019).

In the initial annotation of the corpus we concentrated on a rich set of Named Entities, on some general concepts and events that happen to be frequent in the texts from various genres and domains such as history and ethnography. Handling the typical NEs, events and roles, we plan to support the extraction of knowledge about real people, places, organizations, events etc. and linking them to related facts. The result would be Bulgaria-centered linked data.

In order to control and predict the structure of the extracted knowledge, the annotation scheme followed the philosophy of CIDOC-CRM[1] ontology which has been widely used in the area of GLAM (Galleries, Libraries, Archives, and Museums) and Humanities. In addition, we used information from FrameNet[2], and locally adjusted the scheme to our data. For the creation of the corpus we relied on the INCEpTION annotation tool[3] — (Klie et al., 2018). The models that were initially trained are implementations in spaCy, Flair NLP and Hugging Face.

In spite of the fact that the corpus uses nested annotations at the level of NEs (and the level of events and their roles), the reported experiments use as input only the version of flattened NE annotations. The reason is that given the complexity of the domain texts, we decided to train first a standard NER module with SOTA methods. Thus, our future work is envisaged to handle nested NEs as well as events with their roles. Last but not least, a co-reference model will be trained on the corpus.

In our view the novelties in this paper are the following ones: providing an annotated corpus with NEs, events and roles for Bulgarian in the field of SSH; providing insights related to the annotation scheme and the process of annotation with respect to SSH; using training models as means of detecting inconsistencies in the annotation process and scheme; reporting initial results on the utility of the event corpus on the level of a rich tagset of NEs.

The structure of the paper is as follows: in the next section a focused overview of related works is pre-

[1] https://www.cidoc-crm.org/

[2] https://framenet.icsi.berkeley.edu/fndrupal/

[3] https://inception-project.github.io/

sented. In Section 3 the main ideas behind the annotation scheme are outlined together with some challenges. Section 4 introduces the results from the initial experiments with the event corpus. The last section concludes the paper.

2. Related Work

In this section several lines of related work are presented. Thus, the text below only gives the main tendencies and is not meant to be exhaustive. These lines of related works are as follows: knowledge graphs; event extraction; annotation schemes for event corpora. It should be noted that our main focus here is on the annotation schemes.

Knowledge graphs

We follow the ideas of (Rospocher et al., 2016) on using NLP technologies for the construction of Event-centric knowledge graphs. The authors defined Event-centric knowledge graphs as: *"a Knowledge Graph in which all information is related to events through which the knowledge in the graph obtains a temporal dimension."* This conception is very appropriate to our goal where we expect the knowledge related to Bulgaria to be temporally ordered. In our work we aim at the creation of a similar ontology model and related language pipeline for Bulgarian. One main difference in our approach is the selection of a domain ontology for the type of texts we would like to process.

Event extraction

We are aware that nowadays the event extraction is applied at a large scale and for big quantities of data. For example (Yuan et al., 2018) suggest a fully automatic unsupervised method (Nonparametric Bayesian Model) for extracting event profiles across documents of open-domain news corpora. However, we focus on specific domains. The authors comment that for event extraction two main streams compete - ACE-oriented and MUC-oriented. The former uses triggers like predicates but they are too fine-grained while the latter uses too general event types. Our approach aims at balancing both of them as much as possible.

In contrast to the fully automatic approaches for extracting events, we first rely on a pre-defined annotation scheme and on an annotated event corpus for enhancing trained automatic modules on these data.

Annotation schemes

One of the most popular annotation schemes for event annotation is ACE.[4] ACE defines an approach in which the following elements are annotated: Events, Triggers of the Events, Participants in the Event, and Named Entities. Each Event is annotated with respect to the span of one sentence. The trigger of an event needs to be located within the sentence. Following this line, only participants within the selected sentence are annotated. The similarities of our scheme with ACE are as follows: a standard NE typology is used with some subtypes in addition; nested NEs are considered; coreference is respected although ACE postulates it within the sentence while in our scheme it goes beyond sentence level. Our guidelines differ from ACE in the following aspects: we divide the annotation in two levels - NEs and events with a set of roles. These two levels may overlap and be part of each other. Then, the regular polysemy is handled differently. We extend our NE level with labels like PER-GPE for handling nationalities and regionalities (Bulgarians, etc.) and take decisions locally for the NE as a LOC or ORG, while in ACE the label GPE.PER is used for resolving regular polysemy in context (France vacations in August.) - it is an example of the existence of typed labels there; in ACE the NE subtypes are more detailed in comparison to our scheme. We moved this level of detailness into the next level of events and roles.

There exist a number of more specific annotations schemes. For example, they might be oriented to better handling temporal relations like in (Ning et al., 2018) or focusing on causality and temporality like in (Mostafazadeh et al., 2016). Other schemes divide the annotation into two layers - syntactic (functional) and semantic (ontological) like in (di Buono et al., 2017). Our scheme is in these lines but at the moment it combines the functional and ontological information rather than dividing it into two separate steps.

As it was mentioned above, our annotation scheme allows nested NEs. However, no detailed rules were formulated apart from the need of handling the distinct names within a lengthy name. There was also some inconsistency on whether to include titles within the names (the advantage of this being the disambiguation) or to leave them out. The annotators had the freedom to apply both strategies. Some best practices to follow in our guidelines re-design and specialization are: the work of (Ringland et al., 2019) where consistent substructures are introduced together with the inclusion of roles and categories within NEs; the work of (Plank et al., 2021) where an annotation scheme is presented for nested NEs and experiments are performed on a multi-domain corpus of Danish.

3. The Annotation Scheme and Annotation Process

The aim of the annotation with events and named entities is the inclusion of these data into the Bulgarian-centric Knowledge Graph. Presently it has been done mainly on the basis of knowledge extraction from scientific publications, encyclopedic sources (like Wikipedia), existing structured data. Thus, we consider text as main source of information for the represented objects. At the beginning we focus preferably on the following entities:

- **People** – their biographies – their characteristics, motivations, opinions, events in their lives, roles they played

[4] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

Label	Description
DOC	Various texts, including documents, excluding juridical documents – see **JUR**
EVT	Named events like Second World Wars
JUR	Juridical documents: laws, regulations, etc.
LOC	Locations/places — natural or man-made like mountains, lakes, etc., geopolitical units are excluded – see **LOC-GPE**
LOC-GPE	Geopolitical units (countries, regions, cities, cantons, etc.)
MSC	Miscellaneous names that not included in the other categories
MSR	Measurements with expressed quantity
ORG	Organizations of any kind
PER	People (existing in reality or fictional ones)
PER-GPE	Nationalities (Bulgarian), the birth place, or the place where people live
PER-GRP	Groups of people that cannot be described as **PER-GPE** or **PER-LOC** (Slavs, etc.)
PER-LOC	People that are related to geographical region, but not **PER-GPE**
PRO	Products — tangible and intangible (**DOC** and **JUR** excluded)
REF	Bibliographical references, citations of them, links.
SUM	Amounts of money — a subclass of **MSR**
TIME	Time points or periods

Table 1: Named Entity labels in the annotation scheme.

- **Organizations** – their establishment, life cycle, activities, etc.
- **Objects** – geographical, artefacts, etc. and their features
- **Events** – place, time, participants (People, Objects), relations to other events
- **Time and Periods** – ordering of events in time
- **Documents** – authors, content, opinions, mentions of people, events, entities, etc.

Our annotation scheme reflects the rationale behind the CIDOC-CRM ontology since this ontology has been widely used in the areas of GLAM and Humanities. The annotation scheme envisages two main layers: the first one is the Named Entity (NE) layer, and the second one is the event layer where each event is connected to its participants. The annotation process followed this differentiation. At the first stage, only the NEs were annotated. The events and their participants were annotated at the second stage. Thus, the latter process relied on the already annotated NEs. We also annotated the co-referential links among participants within events across the texts. This is necessary in order to support the extraction of facts from texts even when the participants are not explicitly mentioned within the text span of the event.

In contrast to the ACE guidelines, our annotation scheme does not rely so much on triggers for specific events because in many cases a clear trigger is not present within the text or it might be prone to ambiguity. As a consequence, the event depends on the annotated span, the genre of the text and its typical strategies of text organization. For example, in biographies the dates of birth and death are given in brackets without explicit predicates. For that reason we decided to exploit the so-called *span based annotation* within which the participants are annotated. For more details see (Laskova et al., 2020).

The set of named entity categories covered the main ontological types like persons, locations, organizations, and time. Gradually, the types became more specialized depending on the specifics of the domain text thus resulting in 16 types altogether. Some of these are specializations of the main ones and other are specific for a certain domain/genre. There are specializations of Person and Location. The specific ones include **JUR**(idical) for legal documents, **REF**(erence) for bibliographical sources, among others. It is worth mentioning that already at this level events are annotated as **EVT** (for example, wars, sports events, etc.). Three very similar classes for people are **PER-GPE**, **PER-LOC**, and **PER-GRP**. They reflect some subtle differences in categorization of the group of people. The presence of these differences would be valuable as part of the extracted knowledge. However, with respect to the usage of the corpus itself, it was not clear whether the automatic models could handle well such distinctions. Thus, we decided to group them together in the lable **PERS** — see below the section on experiments. The category **MSR** was extended to cover not just measurements, but also any quantities of objects including people, animals, tools, goods, etc. Unfortunately, the group became very heterogeneous and the automatic methods performed very poorly. For that reason, this category was excluded from the experiments. Table 1 presents the whole set of Named Entity labels in the

Event	Roles
Donation	**donor** (person or organization) **recipient** (person or organization) **theme** (object) **mediator** (person or organization, it could be fund) **period–of–iterations** (time: the length of time from when the event denoted by the target began to be repeated to when it stopped) **goal** (situation: the goal for which the donor gives the theme to the recipient) **time** **place**
Giving–Birth	**brought–into–life** (the new born person) **parents** (the mother and father expressed together, for example "his parents" or "Penka and Toncho Ivanovi") **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)
Moving–in–Place	**agent** (a person) or theme (another type of object) **coagent** (another person or group of people the agent is moving with) **move-from** (the place from which the agent or the time moves) **move-to** (the place where the agent or the theme moves to) **time/beginning/end/duration** **purpose** (a situation or another event which causes the moving) **goal** (a situation/event to be achieved with the moving)
Leaving	**agent** a person or an organization that leaves a group **group** the group of which the person or organization ceases to be a member **time** the moment when the event leaving is performed **reason** why the leaving was performed
Characterisation	**characterised** (a person, organization, etc.) **characteristic** **evaluator** (the one who points out the characteristic: "Gorbachev insists on "second perestroika", Russia considers him Judas.") **source** (a document containing the characteristic)
. . .	. . .

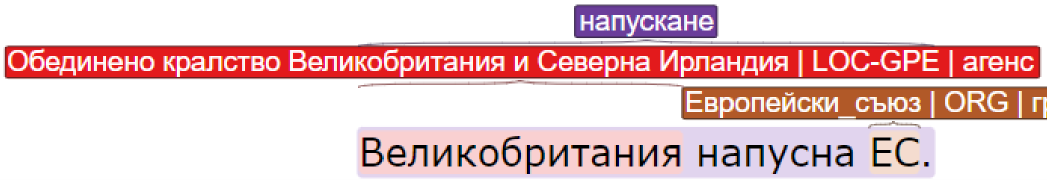
Table 2: Some of the event labels in the annotation scheme.

current scheme.

Since the event types coming from CIDOC-CRM (like End of Existence, Death, Activity, Modification) are too general to cover all specificities of the domains, the annotation scheme was extended on the basis of the English Framenet. For each event a set of participants were defined and mapped to the appropriate places in the CIDOC-CRM hierarchy of events. The formal relations between the ontology and the Framenet-based event schemes are planned to be established later.

In the current annotation scheme some of the events are left general while other are detailed according to the text needs. The reason for this is that the scheme was developed incrementally and empirically with more data being added in various SSH domains. Examples for general events (higher in the hierarchy) are causation, has-parts, start, end, possession, change, existence, destruction. Examples for specific events are rent, teaching, occupation, publication, making a copy, charity. Each event is augmented with its typical roles. The Table 2 presents some of the event labels together with the appropriate participant roles.

In the examples within the table we can see different levels of detailness of the related participants. In the case of the event **Donation** there is a quite detailed set of roles, including *donor*, *recipient*, and *theme* which are expected to be present or presupposed in each instance of this event. However, the annotator might select also some secondary roles associated with the event like *mediator*, *goal*, and *period-of-iterations*. These are not obligatory. It can be noted that general roles like *time* and *place* are included as well due to their importance for each act of donation. In the case of the event **Giving-Birth** the specific roles of *mother* and *father* are specified, on one hand, and the more general and aggregated role of *parents* is presented, on the other. Which one of them will be used in a given text, depends on the way they are presented — as a col-


напускане
Обединено кралство Великобритания и Северна Ирландия | LOC-GPE | агенс
Европейски_съюз | ORG | група

Великобритания напусна ЕС.

Figure 1: The annotation in INCEpTION of the sentence “Great Britain has left EU.”.

lective body or separately. The event **Moving-in-Place** demonstrates a general case of relocation of some objects or people in space. It might be applied in the cases of: people moving to live in another town or country; migration of people; moving of troops from one position to another; moving some artefacts from one museum to another, etc. Which one from all these cases is featured, depends on the semantics of the text — then on the moving verb within the text as well as the related participants. In the annotation scheme in addition some very general types of events were included like **Event** (not present in the table) which are used in cases when there is an important event mentioned in the text, but it does not fit the available events in the scheme. Another case of such general event is the so-called **Characterization** which was included to describe a given feature typical to an object. These two general classes are some of the most frequent within the corpus. The idea is to observe the cases within these general types through concordances and further classify them into more specific sub-events.

The corpus comprises a wide variety of domain texts: historical texts from different periods of Bulgarian history; cultural artefacts like icons; scientific publications; archival documents; encyclopedic articles from Bulgarian Wikipedia. The selection of the such a diverse set of domain texts was done with the aim to develop an annotation scheme that would be applicable for all of them.

We have included texts on important people, places and events from the Bulgarian history. Their importance was determined in two ways. First, we consulted experts from different research institutions who provided the names of these entities, but also some texts about them. The second approach was through selecting the names of streets in about 30 largest cities in Bulgaria, assuming that the most frequent names reflect their importance in our history. Concerning people, through linking we include not only the names related to the most popular ones, but also the same names held by other people. For example, the name Petar Stoyanov refers to a previous president of Bulgaria, but also to a Bulgarian sumo fighter, Bulgarian footballer, Bulgarian historical persons, etc. Also for some people we include other entities named after them like towns, villages, squares, stadiums, sport teams, etc. The texts were cleaned from formatting (we kept the links in Wikipedia articles for later use), then they were segmented into sentences and loaded into INCEpTION system for the annotation. Each document was annotated by two annotators. Then the documents were curated by a superannotator.

The Kappa inter-annotator agreement was measured on the named entities and the events within INCEpTION only on the biographical subset of documents. The values for named entities are between 0.87 and 1.0 for the different types of NEs. The values for events are between 0.87 and 0.91 for the different types of events. The figures reflect the expectations that the event annotation is a harder task. It should be noted also that inside the NEs annotation some categories are harder than others. For example names of documents and organizations are more complex than names of people and locations. Another source of disagreement stems from the guidelines where some freedom has been granted in the selection of shorter or longer spans. This fact inevitably influences the evaluation. In the event annotation there is a pre-defined hierarchy of events and when in doubt, the annotators can select more the general type instead of the recommended most specific one.

As mentioned above, in addition to Named Entities and Events layers, we also annotated co-reference chains in cases when one element of the chain is a role within some of the events in the text. The established coreference chains allow us to extract facts about NEs even when they are not mentioned explicitly within some sentences. In the next step of corpus annotation we envisage linking of each named entity with the appropriate URI from a knowledge base — this process had already started, but it is still far from being completed. We will exploit the co-reference chains for distributing the URIs to other places within the documents.

Here a small example is included of the process of annotation exploitation of the annotation for extracting facts from the texts. In Fig. 1 the annotation of the sentence “Great Britain has left EU.” “Great Britain” is annotated as **LOC-GPE**, “EU” is annotated as **ORG**. The whole sentence is annotated as the event **Leaving** which has four roles (see the Tab. 2). Two of them are present in the sentence — “Great Britain” is playing the role of the *agent* of the event and “EU is the *group*. Additionally, “Great Britain” and “EU” are mapped to their corresponding URIs in the Bulgarian DBpedia.[5]

[5] In the DBpedia dump we use currently, for the annotation there are no facts about BREXIT. But in the Bulgarian

The mapping between the annotation scheme and the ontology CIDOC-CRM supports the extraction of the facts from this annotation. Here are the relevant excerpts from the CIDOC-CRM ontology:

Class: **E86 Leaving**
Subclass of: **E7 Activity**
Scope note: This class comprises the activities that result in an instance of **E39 Actor** to be disassociated from an instance of **E74 Group**. This class does not imply initiative by either party. It may be the initiative of a third party.
In the Annotation Scheme it corresponds to the event label **Leaving**.

Class: **E39 Actor**
Subclass of: **E77 Persistent Item**
Superclass of: **E21 Person** and **E74 Group**
Scope note: This class comprises people, either individually or in groups, who have the potential to perform intentional actions of kinds for which someone may be held responsible.
In the Annotation Scheme it corresponds to NEs labels: **PER**, **ORG**, and **LOC-GPE** due to the regular polysemy of these kinds of Named Entities.

Property: **P145 separated** (left by)
Domain: **E86 Leaving** Range: **E39 Actor**
Subproperty of:
E5 Event **P11 had participant** *E39 Actor*
Quantification: many to many, necessary $(1, n : 0, n)$
Scope note: This property identifies the instance of **E39 Actor** that leaves an instance of **E74 Group** through an instance of **E86 Leaving**.
In the Annotation Scheme it corresponds to the role label: **agent**

Property: **P146 separated from** (lost member by)
Domain: **E86 Leaving** Range: **E74 Group**
Subproperty of:
E5 Event **P11 had participant** *E39 Actor*
Quantification: many to many, necessary $(1, n : 0, n)$
Scope note: This property identifies the instance of **E74 Group** an instance of **E39 Actor** leaves through an instance of **E86 Leaving**.
In the Annotation Scheme it corresponds to the role label: **group**.

By means of these mappings together with the annotation we can extract the following RDF statements:

```
dbpedia:Great_Britain a
        cidoc:E74_Group .
dbpedia:Great_Britain a
        cidoc:E74_Group .
dbpedia:European_Union a
        cidoc:E74_Group .
dbpedia:Brexit a cidoc:E86_Leaving .
dbpedia:Brexit cidoc:P145_separated
        dbpedia:Great_Britain .
dbpedia:Brexit cidoc:P146_separated_from
        dbpedia:European_Union .
```

The RDF statements extracted in this way from different texts will be loaded into a RDF repository for further processing. Also beside the factual information from the text we need to store provenance information about the documents, the processing steps, etc.

4. Experimental Settings

In this section we describe the settings of the first experiments with the corpus — Named-entity recognition. The event processing requires much more examination and thus we leave it for future work.

For training the flat NER model, we divided our dataset of 325 annotated files randomly into three sets - training (262 documents/11803 sentences), development (33 documents/1423 sentences) and test (32 documents/2124 sentences).

For the experimental work we compare Spacy's [6] Transition-based approach to NER with Flair[7], an NLP library implemented on top of PyTorch[8].

For a baseline model, we use the build-in spaCy named entity recognizer, based on transition-based parsing (TBP) and Bloom embeddings (Serrà and Karatzoglou, 2017) that give a good balance of efficiency and accuracy. We train the model with default parameters, except for a larger batch size (*batch_size = 50000*) for speeding.

Flair, on the other hand, provides easy python interfaced access to their own pre-trained Flair contextualized pooled embeddings (Akbik et al., 2019) and many other state-of-the-art language models, such as FastText (Grave et al., 2018), GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018) and the Transformers provided by HuggingFace[9]: BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet. Stacking the embeddings is one of the most important features of the library and the functionality is used in the experiments to concatenate language models together. The developers of the library claim that this method often gives the best results and lately has become a common technique in sequence labeling models (Grave et al., 2018).

We use two types of embeddings stacked in the Flair model. Byte pair Embeddings (Heinzerling and Strube, 2017) use subword information to try to solve the unknown words problem. This feature is important in our case since the dataset consists of historical texts in Bulgarian and the language changed significantly in the last 100 years. There are a lot of words and word forms that are no longer used in modern language, thus the problem with unknown words in common word embeddings fits exactly our data. For the same reason the second embeddings that we stack together with the byte

Wikipedia there are such facts, and we will use it in this example.

[6] https://spacy.io/

[7] https://github.com/flairNLP/flair

[8] https://pytorch.org/

[9] https://github.com/huggingface

pair embedding are character based ones (Santos and Guimaraes, 2015) proven to boost the results on NER for morphologically rich languages like Bulgarian.

5. Results and Evaluation

For a baseline the NER-TBP model was used. It was trained with the standard spaCy parameters except that the batch size was changed, as mentioned above.

Not surprisingly, the best detected labels are LOC-GPE (F1 87.72) and PER (F1 85.11). The next well detected labels are PERS (F1 79.20) and EVT (72.27).

When comparing with our previous work on NER over contemporary newsmedia data (see (Marinova, 2019)), it can be seen that here the F1 values are lower and thus far from NER SOTA. However, as already discussed above, here we train models on a multi-domain corpus where different varieties of language are used, different name models are used and where the NE annotation is richer. From this perspective, the reported here initial results are promising for some labels (like PER, LOC-GPE, PERS, EVT) and need re-considering of our strategies to other labels (like DOC and PRO, also JUR, LOC, ORG).

As we can see in Table 3 there are three labels that underscore the average model performance: **JUR**, **DOC**, and **PRO**. This is, on the one hand, due to the limited number of examples for them, but on the other, we observe that the above mentioned labels overlap semantically in the annotation scheme and vary in the annotated texts. It should be further specified in our guidelines in order to achieve better inter annotator agreement.

It is obvious that the BiLSTM model gives the best results of F1 for all the labels except for the labels **TIME** and **DOC** where the baseline performs better. For the former label we assume that the result is such due to the better performance of the token-based embedding approach in the baseline on numbers in comparison to the character-based embedding one in BiLSTM. As mentioned above, on **DOC** both models perform poorly.

6. Qualitative Error Analysis

The main errors spotted during the quantitative analysis can be summarized as expected in the following groups: model-based and annotation-based. In the examples below the first column represents the gold annotation and the second one refers to the annotation made by the model.

The observed discrepancies between the moderately cleaned gold data and the model are due to the following factors:

- *The existence of different tags for similar entities in the annotation scheme.* Here the model might choose the alternative. A typical case is the following: the film and book titles are erroneously annotated in the data as **PRO** in analogy to song and play titles but the model tags them correctly as **DOC**, because the latter label refers to text documents with the exception of the juridical ones (**JUR**). Or vice versa, the gold data present the correct analysis with **DOC** while the model predicts **PRO**. See an example for the latter:

  Пространните [Extended] B-DOC B-PRO
  жития [biographies] I-DOC I-PRO
  на [of] I-DOC I-PRO
  Кирил [Cyril] I-DOC I-PRO
  и [and] I-DOC I-PRO
  Методий [Methodius] I-DOC I-PRO

- *The overgeneration nature of the annotation.* The allowance of the longest span might cause problems because when a date is taken as a modifier to a noun phrase, the whole phrase is annotated as **DOC**, while the model recognizes the annotation only of the date as **TIME**. For example, see the next excerpt:

  в [in] O O
  цитирания [cited] B-DOC O
  дипломатическия [diplomatic] I-DOC O
  доклад [report] I-DOC O
  от [from] I-DOC O
  1978 I-DOC B-TIME
  г [year] I-DOC I-TIME
  . I-DOC I-TIME

  Another example is a location (**LOC**) that modifies an event (**EVT**). In the gold annotation all the elements are given as **EVT** while the model treats the modifying prepositional phrase as **LOC**:

  на [at] O O
  Световното [the World] B-EVT B-EVT
  първенство [championship] I-EVT I-EVT
  в [in] I-EVT O
  САЩ [the USA] I-EVT B-LOC

- *The model might not recognize correctly the specific context.* In the first example, given below, the name 'Bulgarians' is a name of a village as indicated by the pre-positioned nominal classifier 'village'. However, the model treated is as a nationality name. In the second example, the person name is used as a name of a city, but again it is treated by the model as the original one - **PER**.

  село [village] O O
  Българи [Bulgarians] B-LOC B-PERS

  гр. [city] O O
  Гоце [Gotse] B-LOC B-PER
  Делчев [Delchev] I-LOC I-PER

- *Handling regular polysemy.* In the example below the General Assembly is annotated as **EVT** but the model treats it as an **ORG**. Both annotations are valid. The **PER** is correctly identified:

		NER-TBP			BiLSTM		
Entity	**Examples**	**P**	**R**	**F1**	**P**	**R**	**F1**
TIME	495	80.87	71.51	**75.90**	70.00	81.01	74.81
LOC	197	68.53	58.95	63.38	71.60	61.42	**66.12**
LOC-GPE	641	80.99	84.43	82.67	84.32	91.42	**87.72**
PER	858	79.61	83.05	81.29	85.26	84.97	**85.11**
ORG	300	64.47	67.77	66.08	67.55	68.00	**67.77**
JUR	17	34.09	36.59	35.29	53.85	41.18	**46.67**
EVT	126	68.60	50.43	58.13	76.79	68.25	**72.27**
PERS	134	82.46	59.49	69.12	85.34	73.88	**79.20**
DOC	35	29.79	24.56	**26.92**	23.08	25.71	24.32
PRO	112	24.24	12.50	16.49	36.51	20.54	**26.29**
SUM	195	50.00	10.34	17.14	65.78	63.08	**64.40**
All	3114	74.85	71.38	73.08	76.29	76.69	**76.49**

Table 3: Results for the named entity recognition task on the Bulgarian Event Corpus. We compare two algorithms - transition-based parsing for named entity recognition (NER-TBP) and Flair BiLSTM with stacked byte pair and character embeddings (BiLSTM)

След [After] O O
закриване [closing] O O
на [of] O O
Общото [General] B-EVT B-ORG
събрание [Assembly] I-EVT I-ORG
Стефан [Stefan] B-PER B-PER
Стамболов [Stambolov] I-PER I-PER

The presented error analysis shows the problematic places in the annotation scheme and the annotation process with respect to certain NE labels. Also, it raises some issues about the proper handling of nested NEs. Thus, it will be very useful in the next specifications of the annotation model.

7. Conclusions and future work

The main contribution of this paper is the enrichment of the Bulgarian language resources set with a multi-domain event oriented dataset with a clearly involved diachronic dimension. Also, some initial experiments have been performed on the NER task over this dataset. Standard NER task is not trivial on a corpus with multiple domains in SSH and also with complex NER and event annotations. The result are promising for some NE labels and show the weaknesses of our annotation approach to other set of labels. In this way, we use these experiments as a means for error detection and insight trigger for the next improvements of the annotation scheme and annotation process. The BEC will be made available through ELRA and through CLaDA-BG repository. Within CLARIN it can enhance also a Resource Family of Event corpora for various languages.

Our future work will concentrate on the formulation of stricter annotation rules for NER labeling and nesting in order to simplify the task for the human annotators and thus achieve better results later on the underscoring labels. Much more example data is necessary to ensure enough material with these now scarce labels. Training modules on our current (although imperfect) data will also help us produce more data in an automatic way and post-edit it accordingly.

From an experimental point of view, in our next steps we intent to explore some of the state-of-the-art models for Nested Named Entity Recognition, Event Extraction and Relation Extraction in order to support further the creation and enrichment of the Bulgarian Event Corpus (BEC) as well as the Bulgarian-centric knowledge graph.

8. Acknowledgements

The reported work has been partially 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, Grant number DO01- 377/18.12.2020.

We would like to thank the anonymous reviewers for their valuable comments. All errors remain our own. Also, we would like to thank the annotators Iva Anastasova, Lyudmila Dimitrova, Preslava Georgieva, Nikola Ikonomov and Laska Laskova for their work.

9. Bibliographical References

Akbik, A., Bergmann, T., and Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728.

di Buono, M. P., Tutek, M., Šnajder, J., Glavaš, G., Dalbelo Bašić, B., and Milić-Frayling, N. (2017). Two layers of annotation for representing event mentions in news stories. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 82–90, Valencia, Spain, April. Association for Computational Linguistics.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Heinzerling, B. and Strube, M. (2017). Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. *arXiv preprint arXiv:1710.02187*.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Laskova, L., Osenova, P., and Simov, K. (2020). Towards an interdisciplinary annotation framework: Combining nlp and expertise in humanities. In *Proceedings of CLARIN Annual Conference 2020*.

Marinova, I. (2019). Evaluation of stacked embeddings for Bulgarian on the downstream tasks POS and NERC. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 48–54, Varna, Bulgaria, September. INCOMA Ltd.

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., and Vanderwende, L. (2016). CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California, June. Association for Computational Linguistics.

Ning, Q., Wu, H., and Roth, D. (2018). A multi-axis annotation scheme for event temporal relations.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Plank, B., Jensen, K. N., and van der Goot, R. (2021). Dan+: Danish nested named entities and lexical normalization.

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., and Curran, J. R. (2019). NNE: A dataset for nested named entity recognition in English newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181, Florence, Italy, July. Association for Computational Linguistics.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37-38:132–151.

Santos, C. N. d. and Guimaraes, V. (2015). Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*.

Serrà, J. and Karatzoglou, A. (2017). Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 279–287.

Simov, K. and Osenova, P. (2019). Integrated language and knowledge resources for clada-bg. In *Selected Papers from the CLARIN Annual Conference 2019*, pages 137–144. Linköping Electronic Conference Proceedings.

Yuan, Q., Ren, X., He, W., Zhang, C., Geng, X., Huang, L., Ji, H., Lin, C.-Y., and Han, J. (2018). Open-schema event profiling for massive news corpora. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 587–596, New York, NY, USA. Association for Computing Machinery.