The Bulgarian Event Corpus (BEC): Models for Extracting Knowledge from Text

Petya Osenova, Kiril Simov, Nikolay Paev, Stefan Marinov,
Kristin Dimitrova, Tsvetelina Aleksandrova
Institute of Information and Communication Technologies, BAS

1. Overview

- 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 Bulgaria-centric Knowledge Graph
- The corpus comprises a wide variety of domain texts:
    - historical texts from different periods of Bulgarian history;
    - cultural artefacts like church icons;
    - scientific publications;
    - archival documents;
    - encyclopedic articles from Bulgarian Wikipedia
- In the initial annotation of the corpus we concentrated on:
    - a rich set of Named Entities,
    - some general concepts and events that happen to be frequent

2. BEC Specifics and Annotation Schema

- In order to control and predict the structure of the extracted knowledge, the annotation scheme followed the philosophy of CIDOC-CRM - ontology which has been widely used in the area of GLAM (Galleries, Libraries, Archives, and Museums) and Humanities
- In addition, we used event descriptions from FrameNet, and locally adjusted the scheme to our data.

Token_id	Token	Ne_1	Ne_2	Ev_1	Ro_11	Ro_12
29	Той	C<		з<	а<	
30	е			а	т<	
31	министър-председател			н	р	п<
32	на			и		
33	България	L<		м	в<	
34	в			а		
35	правителството	O<		н	в<	
36	на	R		и	ъ	
37	БЗНС	G	O<	е	з	
38	(				л	
39	1919		T<		о	
40	-				ж	
41	1923		T<		и	
42	)				т	
43	.					

Елемент	Референция	URL	Текст	F_token	L_token	To check
КОМЕНТАР						
▼ НАИМЕНОВА						
COREF		Александър	Той	29	29	T
LOC-GPE		България	България	33	33	F
ORG		Правителств	правителств	35	42	F
ORG	Български з	Български_з	БЗНС	37	37	F
TIME	1919 г.	dummy:1919	1919	39	39	F
TIME	1923 г.	dummy:1923	1923	41	41	F
▼ СЪБИТИЯ						
▼ занимани			Той е минис	29	42	F
агенс	Александър	Александър	Той	29	29	F
тригер			е министър-	30	31	F
позици		министър-пр	министър-пр	31	31	F
възлож		България	България	33	33	F
възлож		Правителств	правителств	35	42	F

- Several Annotation Levels: Named Entity Annotation - PER, LOC, LOC-GPE, ORG, TIME; Events and Roles – Event definition and participants roles: Event(Role01, Role02, …). Kappa: between 0.87 and 1.0
- Events are selected on the basis of ontology CIDOC-CRM and extended with frames from FrameNet: between 0.87 and 0.91
- Named Entities are annotated with identifiers from Bulgarian DBPedia

3. Named Entities Types

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
COREF	Words that co-refer to a named entity (e.g. pronouns) in anaphoric co-referential chain

4. Event Types

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)

5. The Experimental Set Up

We use our own pre-trained general purpose LLMs.

- For the NER we use BERT models
- We model the Event Extraction as conditional generation and thus we use encoder-decoder T5 models and decoder-only Llama models

The model generates a list of events from an input sentence.

For the output structure, we consider a structural representation in a JSON compatible scheme. The models were trained to produce a list of dictionary data structures, each representing a single event. Each event has three fields: type, text span and list of roles - with type and text span

```
[{
    "type": "occupation",
    "text": "He is a prime-minister of
Bulgaria in the government of BZNS
(1919 - 1923).",
   "roles": [
      ["agent", "He"],
      ["position", "prime-minister"],
      ["assignor", "the government
           of BZNS"],
      ["beginning", "1919"],
      ["end", "1923"],
      ["assignor", "Bulgaria"],
      ["trigger", "is a prime-minister"]
   ]
}]
```

Example event representation of the sentence: He is a prime-minister of Bulgaria in the government of BZNS (1919 – 1923)

We fine-tune the models to generate a list of JSON objects, conditioned on an input sentence. The training is done for 10 epochs on sentence level with batch size of 256 and linearly decaying learning rate of 3e-04

6. Experiments and Results

Model	Accuracy	Macro F1
BERT-Base	0.9279	0.8026
BERT-Large	0.9305	0.8123

Table 5: NER metrics.

Model	Precision	Recall	F1
T5	0.8673	0.8199	0.8429
Llama	0.6748	0.7688	0.7187

Table 2: Event extraction metrics.

Model	$\overline{\text{IoU}}$	Type accuracy	Role F1
T5	0.9701	0.9025	0.9026
Llama	0.8931	0.7534	0.8075

Table 3: Metrics over the true positives. The column $\overline{\text{IoU}}$ shows the mean Intersection over union of the span predictions. The **Role F1** column shows the F1 metric over the extraction of the role spans.

Results show that encoder-decoder models (like T5) are more suitable for the event extraction task.