NerKor+Cars-OntoNotes++

Attila Novák[1,2], Borbála Novák[1,2],

[1]MTA-PPKE Hungarian Language Technology Research Group • Budapest, Hungary
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics • Budapest, Hungary
{surname.firstname}@itk.ppke.hu

ABSTRACT

- we present an upgraded version of the Hungarian NYTK-NerKor named entity corpus, which contains
  - about twice as many annotated spans, and
  - 7 times as many distinct entity types as the original version

} newest and biggest NER corpus for Hungarian containing diverse domains

- we used an extended version of the OntoNotes 5 annotation scheme
- we also trained and released a transformer-based NER tagger for Hungarian

RESOURCES

- many legacy named entity corpora contain an annotation distinguishing four entity types: ORG, PER, LOC, MISC →all existing Hungarian corpora use this schema
- OntoNotes 5: GPE, FAC, PROD, LAW, EVENT, WORK OF ART, dates and times, cardinal and ordinal numbers, quantities, percentages and amounts of money, NORP, LANGUAGE
- biomedical, non-English NER corpora
- token-level annotation vs nested entities

corpus	granul.	non-ent	der/cmpd	nest	cmn N	subtok	language	domain	gold
CONLL 2002/2003	-						en, ne, es	news	+
Szeged NER	-						hu	business	+
Criminal NER	-						hu	news	+
hunNERwiki	-						hu	wiki	-
NYTK-NerKor	-						hu	multi	+
OntoNotes 5	+	+					en	multi	+
Genia	-			+			en	biomed	+
NoSta-D	-		+	+			de	wiki+news	+
DaN+	-		+				dk	multi	+
AnCora	-	+		+			es, ca	news	+
CNEC 2	++	+		+			cz	?	+
ACE datasets	++	+		+	+		en, ar, zh, es	news	+
NNE	+++			+			en	news	+
NEMO	+		+	+		+	he	news	+

THE NERKOR CORPUS

- size of over one million tokens,
- has a broad coverage of domains and topics
- consists of five 200,000-token subcorpora: *fiction, legal, web, news, wiki*
- one third of the corpus contains coherent text, the rest is a shuffled collection of unrelated sentences or sentence fragments,
- 20% of the corpus includes morphological annotation as well

ANNOTATION METHOD

Zero-shot Preannotation: we applied two models trained on the English OntoNotes 5 corpus to the Hungarian corpus

1. created by the DeepPavlov team fine-tuning multilingual BERT
2. based on XLM-RoBERTa

Error Analysis and Automatic Error Correction using regular-expression-based patterns → Benefits and dilemmas

e.g. definite article, numerical and quantity expressions

e.g. - NORP vs. GPE/ORG
- time expressions and quantities;
- annotation of adjectives corresponding to English prepositional phrases

3. NameTag 2: Czech model of the NameTag 2 neural named entity tagger trained on the Czech Named Entity Corpus CNEC 2 →fine-grained hierarchy of entity classes having many subclasses within the broader categories

A lemmatized named entity list and automated correction patterns

↓ all alternative analyses for each entity along with their corpus frequencies →identify elements frequently misclassified in the OntoNotes model, and entities that should be assigned to distinct classes (e.g. MEDIA, SMEDIA)

Manual error correction

- corrected anomalies in the lemmatized named entity list,
- resolved contradictions of the annotations in the original corpus and those generated by the transfer models
- normalized the annotation for references to legislation in the law subcorpus

THE CORPUS

FEATURES OF THE CORPUS

The number of distinguished entity types increased 7-fold while the number of entities marked almost doubled

PER[a]	15266	PER[a]	15234	LOC	2226	FAC	831	PROJ	254
LOC	12988	GPE[b]	13872	WORK_OF_ART	1975	MONEY	681	MISC	117
ORG[a]	12343	DATE	11224	QUANTITY	1918	TIME	661	ID	83
MISC	5751	ORG[a]	9512	CAR	1423	EVENT	627	AWARD	64
		CARDINAL	6710	DUR	1395	LANGUAGE	499		
NYTK-NerKor		NORP	4551	PERCENT	1257	AGE	336	**NerKor+Cars**	
tokens	**entities**	ORDINAL	3258	PROD	1174	MISC-ORG	306	**tokens**	**entities**
1027218	**46348**	LAW	3245	MEDIA	1062	SMEDIA	271	**1038947**	**84831**

non-entities	*DATE*	dates and intervals (granularity over 24 hours)
	CARDINAL	cardinal numbers
	NORP	nationalities, religion, political affiliation (adjectives)
	ORDINAL	ordinal numbers
	LAW	references to laws, directives and other norms
	QUANTITY	quantites: cardinal number + unit of measure
	DUR	time durations (time quantites, unanchored to the timeline)
	PERCENT	percentages and ratios (in OntoNotes: only percentages)
	TIME	time and short intervals (granularity below 24 hours)
	LANGUAGE	names of languages
	AGE	age of persons and things (time durations with spec. semantics)
	MONEY	sums of money: cardinal number + monetary unit
organizations	ORG	organizations: companies, parties, institutions, teams etc.
persons	PER	people, fictive persons, families, animals
places	GPE	geopolitical entities: states, settlements, provinces, counties etc.
	LOC	geological locations: mountains, deserts, bodies of water etc.
	FAC	facilities: roads, streets, buildings etc.
other entities	WORK_OF_ART	titles of creative works
	PROD	products (except motor vehicles)
	MEDIA	journals, tv channels, news sites
	CAR	motor vehicles
	SMEDIA	social media
	EVENT	named events (except projects)
	PROJ	projects and programmes
	AWARD	awards
	MISC-ORG	organization-like types of residual entities
	MISC	residual entities

MODELS AND PERFORMANCE

version	original			Det fixed			only labels in common			com. labels, Det fixed		
model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
CZ	15.82	11.39	13.25	15.89	11.44	13.30	64.57	52.92	58.16	64.63	52.97	58.22
DP	66.32	60.41	63.23	71.66	65.27	68.31	68.79	63.42	65.99	74.63	68.81	71.60
FL	74.81	70.73	72.71	80.59	76.19	78.33	77.68	74.34	75.97	**83.90**	**80.29**	**82.06**
NKC	**91.07**	**88.12**	**89.57**				**91.64**	**89.18**	**90.39**			
test	91.92	87.65	89.73									

- the zero-shot performance of the transferbased models
- evaluation with the tagset normalized to the tags present in the original model
- a neural tagger model based on the Hungarian huBERT contextual language model on the training set of the corpus
- performance of the best model trained on NerKor+Cars on each entity type compared to performance a similar model on the original NerKor annotation

NerKor	F_1	NerKor+Cars	F_1
		DATE	**88.85**
		CARDINAL	**83.78**
		NORP	**87.12**
		ORDINAL	**94.67**
		LAW	**82.12**
		QUANTITY	**91.11**
		DUR	**74.67**
		PERCENT	**84.21**
		TIME	**66.67**
		LANGUAGE	**83.33**
		AGE	**100.00**
		MONEY	**87.50**
ORG	88.45	ORG	**93.33**
PER	95.32	PER	**97.11**
		GPE	91.98
LOC	**92.28**	LOC	76.60
		FAC	80.00
		WORK_OF_ART	**90.27**
		PROD	79.37
		MEDIA	**91.53**
		CAR	**92.86**
MISC	81.85	SMEDIA	73.33
		EVENT	72.73
		MISC-ORG	47.06
		PROJ	66.67
		AWARD	**100.00**
		MISC	66.67
	91.02		89.57/**92.05**

Simon, E. and Vad'asz, N. (2021). Introducing NYTKNerKor, a gold standard Hungarian named entity annotated corpus. In Kamil Ekstein, et al., editors, Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings, volume 12848 of Lecture Notes in Computer Science, pages 222–234. Springer.

Nemeskey, D. M. (2021). Introducing huBERT. In Gábor Berend, et al., editors, XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021), pages 3–14, Szeged. Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Weischedel, Ralph and Palmer, Martha and Marcus, Mitchell and Hovy, Eduard and Pradhan, Sameer and Ramshaw, Lance and Xue, Nianwen and Taylor, Ann and Kaufman, Jeff and Franchini, Michelle and El-Bachouti, Mohammed and Belvin, Robert and Houston, Ann. (2013). OntoNotes Release 5.0. Linguistic Data Consortium LDC2013T19, ISLRN 151-738-649-048-2.

Sevcikova, M., Zabokrtsky, Z., and Kruza, O. (2007). Named entities in Czech: annotating data and developing NE tagger. In Vaclav Matousek et al., editors, Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue, volume 4629 of Lecture Notes in Computer Science, pages 188–195, Berlin / Heidelberg. Springer.