Czech Named Entity Corpus 2.0

Zdeněk Žabokrtský, Jana Straková, Milan Straka

January 8, 2014

The aim of Named Entity Recognition (NER) is to identify proper names in text and to classify them into predefined categories such as names of persons, geographical names, names of organizations etc. The task of NER is motivated by the needs of Natural Language Processing (NLP) applications such as Information Extraction and Machine Translation. Similarly to most other tasks in NLP, it is advantageous to use annotated data when developing a named entity recognizer, especially for training and evaluation purposes. The presented Czech Named Entity Corpus 2.0 is a major update to the Czech Named Entity Corpus 1.0, a first publicly available corpus providing a large body of manually annotated named entities in Czech sentences, including a fine-grained classification. The corpus is available under the CC BY-NC-SA 3.0 license.[1]

The difference between Czech Named Entity Corpus 2.0 and 1.1 are the following:

- changed the named entity hierarchy
  - overhaul the number entities
    - entities of supertype c were merged into n; in order to acommodate bibliographic entities a new type nb "vol./page/chap./sec./fig. numbers" was added
      - $cs \rightarrow oa$
      - $cn \rightarrow nb$
      - $cb \rightarrow nb$
      - $cp \rightarrow nb$
      - $cr \rightarrow n\_, or$
    - entities of supetype q were moved into n
      - $qc \rightarrow nc$
      - $qo \rightarrow no$
    - low frequent entities of supertype n were removed and some renamed and merged
      - removed nm, nr, nw
      - nc was renamed to ns
      - $np \rightarrow no$
      - $nq \rightarrow n\_$
    - some time entities were removed
      - $tc \rightarrow no$
      - $tp \rightarrow no$
      - $tn \rightarrow nc$
      - $ts \rightarrow nc$

[1] `http://creativecommons.org/licenses/by-nc-sa/3.0/`

  - new entity *me* representing email was added
  - *gp* entity was merged into *g_*
  - *mr* and *mt* were merged into new *ms*
  - *oc* entity was merged into *o_*
  - *pb* entity was merged into *p_*
- new data was added
  - 125 sentences with many addresses and emails were added
  - 3000 sentences containing only a few named entities were added so that the resulting corpus better represents the density of named entities (density of named entities in CNEC 1.1 is too high)

1 Classification

The classification of named entities in Czech is an updated version of classification from [2] and can be seen in Figure 1. Classification classes distribution is shown in Table 1.

2 Data formats

Named entities are saved in formats:

- **plain text** – manual annotations in plain text format
- **simple xml** – simple xml format
- **treex** – xml format from Treex [1] (formerly TectoMT) with morphologic analysis
- **html** – html with highlighted named entities

References

[1] Martin Popel and Zdeněk Žabokrtský. TectoMT: modular NLP framework. In *Proceedings of the 7th international conference on Advances in natural language processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] Ševčíková Magda, Žabokrtský Zdeněk, and Krůza Oldřich. Zpracování pojmenovaných entit v českých textech. Technical report, Ústav formální a aplikované lingvistiky, 2007.

Figure 1: Classification classes

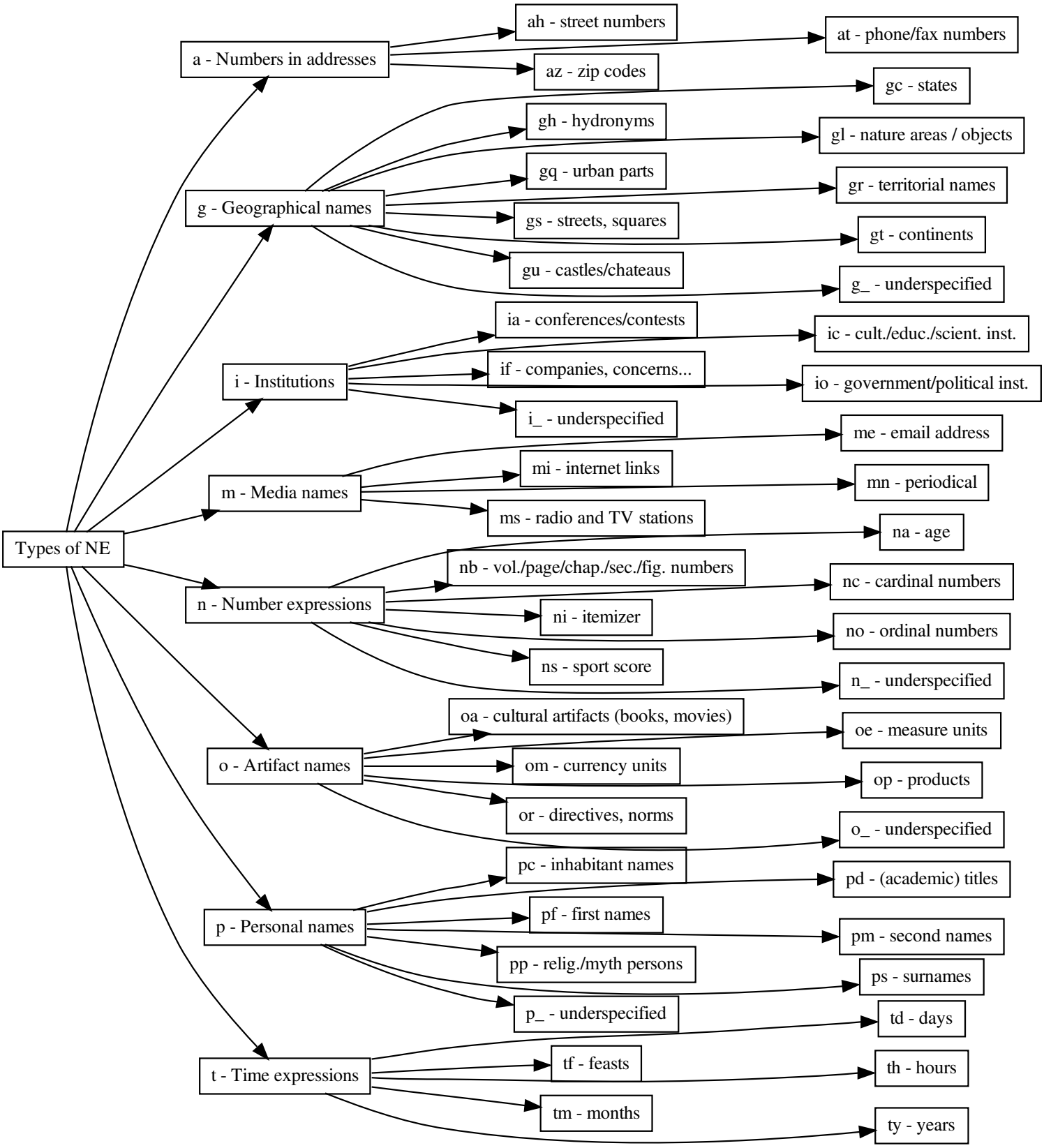

Table 1: Classifaction classes distribution in data

Named Entity	Occurrence		Named Entity	Occurrence	
	Count	Ratio		Count	Ratio
ps	4159	11.81%	mn	241	0.68%
pf	3165	8.99%	cap	240	0.68%
P	2792	7.93%	at	229	0.65%
gu	2718	7.72%	om	210	0.60%
nc	2023	5.74%	gq	200	0.57%
oa	1792	5.09%	ah	190	0.54%
ic	1521	4.32%	ni	175	0.50%
th	1436	4.08%	o_	168	0.48%
ty	1364	3.87%	or	153	0.43%
s	1268	3.60%	A	148	0.42%
gc	1156	3.28%	pp	141	0.40%
if	867	2.46%	ns	137	0.39%
io	802	2.28%	gl	135	0.38%
n_	697	1.98%	az	121	0.34%
tm	584	1.66%	gh	115	0.33%
T	541	1.54%	g_	108	0.31%
f	527	1.50%	na	106	0.30%
td	520	1.48%	me	104	0.30%
oe	476	1.35%	ms	104	0.30%
segm	453	1.29%	pd	101	0.29%
op	409	1.16%	pm	97	0.28%
p_	373	1.06%	gt	96	0.27%
?	358	1.02%	lower	64	0.18%
no	351	1.00%	C	45	0.13%
gs	305	0.87%	i_	40	0.11%
nb	276	0.78%	tf	27	0.08%
pc	258	0.73%	mi	22	0.06%
ia	256	0.73%	upper	6	0.02%
gr	250	0.71%			