Mývalit Topit Baštit . . .

How to Improve Annotation of the Czech Web Corpus?
PV173 Seminář zpracování přirozeného jazyka

Zuzana Nevěřilová

November 22, 2017
Annotation of Czech Web Corpus (2012)

cztenten12 v. 9

annotated by the “Brno pipeline”

- collect data in the Czech language, deduplication
- \texttt{uninorm}: normalize encodings, weird characters etc.
- \texttt{unitok}: Czech tokenizer
- \texttt{desamb}: Czech tagger add lemma and tag
Annotation of Czech Web Corpus (2012)
cztenten12 v. 9

annotated by the “Brno pipeline”

- collect data in the Czech language, deduplication
- uninorm: normalize encodings, weird characters etc.
- unitok: Czech tokenizer
- desamb: Czech tagger add lemma and tag
  - majka: morphological analysis (based on large dictionary of Czech word forms) adds all possible lemmata and tags
  - guessing lemma and tag for unknown tokens (based on context and word ending)
  - disambiguation based on ILP
Annotation of Czech Web Corpus (2012)
cztenten12 v. 9

annotated by the “Brno pipeline”

- collect data in the Czech language, deduplication
- uninorm: normalize encodings, weird characters etc.
- unitok: Czech tokenizer
- desamb: Czech tagger add lemma and tag
  - majka: morphological analysis (based on large dictionary of Czech word forms) adds all possible lemmata and tags
  - guessing lemma and tag for unknown tokens (based on context and word ending)
  - disambiguation based on ILP

the pipeline is not fully suitable for web texts
Possible Annotation Problems

- non-words (AK47, ááááááááááááááááá)
- typos
- foreign words (association)
- names (Aberdeen, Abu Dhabi, Abú Zábí)
- adopted foreign words (clevelandský, stalkinzích)
- components of MWEs (faux pas, play off)
- interlingual homographs (drop, user)
Possible Annotation Problems

- non-words (AK47, ááááááááááááááááááá)
- typos
- foreign words (association)
- names (Aberdeen, Abu Dhabi, Abú Zábí)
- adopted foreign words (clevelandský, stalkinzích)
- components of MWEs (faux pas, play off)
- interlingual homographs (drop, user)

→ use the guesser only on Czech words
The cardio machine from Nová Bašť that was bought by the Maine Coon.
The cardio machine from Nová Bašť that was bought by the Maine Coon.
The cardio machine from Nová Baště that was bought by the Maine Coon.
Baisa Naïve Classifier

“Let \( w \) be a word and \( l \) its respective lemma. If \( l \) does not appear in the corpus then \( w \) is not correctly lemmatized.”

In cztenten12_9:

- 5,069,447,935 tokens
- 68,776,412 token annotations guessed
- 6,432,405 unique lemmata with guessed annotations
- 1,606,599 appear in the corpus as words
- 1,186,041 appear in the corpus as words with freq \( \geq 1 \)
- 552,824 appear in the corpus as words with freq \( \geq 10 \)

nearly 5 million lemmata are probably incorrect
Double Annotation Based on Lists

- MWEs discovered by orthographic instability (a priori, a-priori, apriori) + manually annotated
- Czech place names (all word forms from RUIAN)
- names and surnames
- international place names (Geonames)
- Wikipedia titles
- other (company names, chemical compound names, zoological and botanical names, artwork names, Latin quotations)

overlap with the former categories
The cardio machine from Nová Baště that was bought by the Maine Coon.
Double Annotation Based on Lists

from double annotation to the correct one!

<annotation lemma="Bašť" tag="k1gFnSc3,k1gFnSc6">
Bašti baštit k5eAaImRp2nS
</annotation>

lemmata: baštit (verb), Bašť (noun)
from double annotation to the correct one!

<annotation lemma="Bašť" tag="k1gFnSc3,k1gFnSc6">
Bašti baštit k5eAaImRp2nS
</annotation>

lemmata: baštit (verb), Bašť (noun)

context based?
<annotation lemma="Bašť" tag="k1g FnSc3,k1gFnSc6" type="placeCZ">
Bašti baštit k5eAaImRp2nS
</annotation>
Possible Problems

- over-annotation – choice of the correct annotation
- choice of the correct lemma and tag
- not enough linguistic information (tags for Palo Alto?)
- dictionaries do not solve language mixing (e.g. Zuzanka řekla, že je from Bratislava)
- something unexpected
Future Work

- test the application
- order the resources, improve their quality
- annotate random samples with different settings
- evaluate
- annotate Czech web corpus