Saara: AR on Free Text in Czech

Vašek Němčík
(xnemcik@fi.muni.cz)

NLP Centre
FI MU Brno

RASLAN
December 7, 2012

Outline

Motivation

Example:

(∅) Byl to on.

- we can analyze the structure
- the structure itself is useless
- we need context

The AR task (as in current Saara)

- to identify anaphoric expressions in the given text
  - personal pronouns (including reflexive and zero)
  - *relative pronouns*
  - *demonstrative pronouns*
  - *nominal anaphora*
- to determine the antecedent for each such anaphor
  - only NPs or clauses considered as candidates
- to determine/assume the relation type
  - identity of reference
  - identity of sense

Main principles

- markables (+features)
- anaphoric markables (+type and features)
- consider candidates
  - sorting
  - filtering, ...
- determine links between markables
  - group into coreference classes

Structured vertical format

```
<s id="sent1" class_id="cls1" type="sentence">
<markable id="m1" class_id="cls2" type="clause">
<markable id="m2" class_id="cls3" type="np" gram="subj">
Filip          Filip          k1gMnSc1
</markable>
políbil        políbit        k5eAaPmAgInSrD,k5eAaPmAgMnSrD
<markable id="m3" class_id="cls4" refconstr="m2" type="np" gram="obj">
Lucii          Lucie          k1gFnSc4
</markable>
</markable>
<g/>
.              .              kIx.
</s>
<s id="sent2" class_id="cls5" type="sentence">
<markable id="m4" class_id="cls6" type="clause">
<markable id="ms1" anaref="m2" class_id="cls3" type="pron_pers_zero" gram="subj">
_              on             k3p3gMnSc1,k3p3gInSc1,k3p3gFnSc1,k3p3gNnSc1
</markable>
Miluje milovat k5eAaImIp3nS
<markable id="m5" anaref="m3" class_id="cls4" refconstr="ms1" type="pron_pers_weak" gram="obj">
ji             on             k3p3gFnSc4xP
</markable>
</markable>
<g/>
.              .              kIx.
</s>
```

Morphology + Syntax

`desamb`

- morphological tags
  some need to be dedisambiguated
  (pronouns + finite verb endings)
- sentences

`SET` (or `synt`)

- clauses
- nominal markables
  with projected morphological tags
  pronoun subcategorization

Post-processing of Syntax

- determining grammatical roles
  (heuristics based on cases/prepositions)
- detecting+adding zero subjects
  - no subject in clause
  - *except verb without left valency*
  - *not followed by a "že"-sentence*

Anaphora Resolution

grammatical anaphora + constraints
(based on grammatical roles)

- negative constraints
  based on Chomskyan principles
- resolution of reflexives
  bind reflexives to the clause subject

textual anaphora

- personal pronouns (strong, weak, zero)
- *relative pronouns*
- *demonstrative pronouns*

Saara Demo

`nlp.fi.muni.cz/projekty/anaphora_resolution/saara/demo/`

- input: plain text in Czech
- output: table with markables (and references)

feedback + use appreciated

What am I working on right now?

- de-disambiguation + zero subject detection
- approximating semantic preferences
  using co-occurrence statistics (`wsdump` + `bushbank`)
- translating verticals into instances
  - anaphor features
  - antecedent candidate features
  - combination features
  - `Y/N`

  $\rightsquigarrow$ ML methods
- TFA models

Thank You for Your Attention!