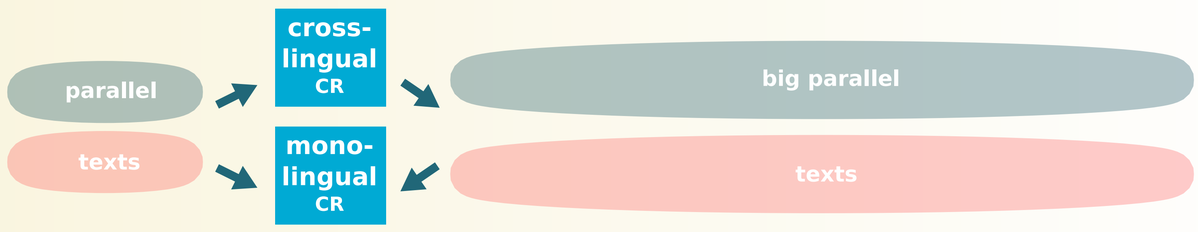
Coreference Resolution System Not Only for Czech

Michal Novák
mnovak@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Charles University in Prague

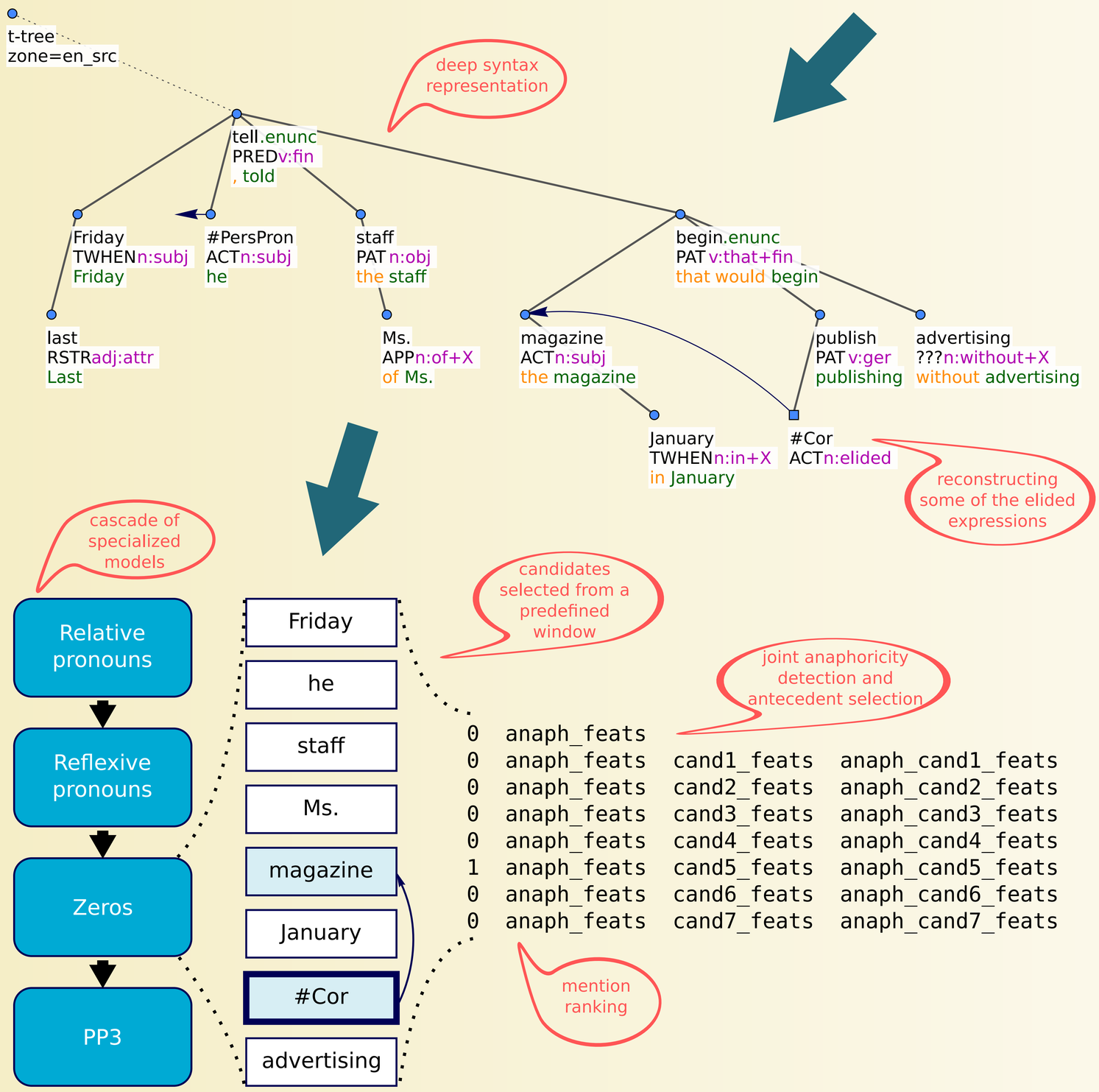
INTRODUCTION

- CR resolver for Czech - the only one to our knowledge
- for English, comparing it with state-of-the-art CR resolvers
- used as a CR system in cross-lingual experiments



TREEX CR ARCHITECTURE

Last Friday, he told the staff of Ms. that the magazine in January would begin publishing without advertising.



PREPROCESSING

- from a plain text up to the level of deep syntax
- using Treex framework (Popel and Žabokrtský, 2010)

- morphology: Morče (Spoustová et al., 2007)
  dependency syntax: MST parser (McDonald et al., 2005)
  named entities: NameTag (Straková et al., 2014)
  transition to deep syntax
  semantic roles: in Treex (Bojar et al., 2016)
  anaphoricity of "it": NADA (Bergsma and Yarowski, 2011)

- morphology: MorphoDiTa (Straková et al., 2014)
  dependency syntax: adapted MST parser (Novák and Žabokrtský, 2007)
  named entities: NameTag (Straková et al., 2014)
  transition to deep syntax
  semantic roles: in Treex (Bojar et al., 2016)

FEATURES

- **Location and distance:**
  anaphor's and candidate's positions
  sentence, clause, word distance
- **(Deep) morpho-syntactic:**
  mention's head POS tag, gender, number, case, semantic role
  parent's features
  valency frame matching
- **Lexical:**
  lemmas of mentions' heads
  noun-verb collocations stats on a large data
  WordNet, EuroWordNet
  named entities

MODEL

- Vowpal Wabbit toolkit
- logistic regression
- optimized by stochastic gradient descent

RESULTS

Stanford Core NLP (Lee et al., 2011; Clark and Manning, 2015; Clark and Manning, 2016)

CR in Treex used for generating CzEng 1.0 (Bojar et al., 2012)

trained and tested on PCEDT 2.0 Coref (Nedoluzhko et al., 2016)

tested on CoNLL 2012 test set (Pradhan et al., 2012)

	PCEDT Eval					CoNLL 2012 test set		
	Relative	Reflexive	PP3	Zeros	All	Reflexive	PP3	All
Count	842	49	2,494	3,260	6,645	111	4,583	4,710
Stanford								
deterministic	1.16	55.67	63.65	0.00	34.96	71.11	60.55	60.79
statistical	0.00	63.74	72.71	0.00	39.09	80.56	*71.07*	**71.29**
neural	0.00	70.97	*76.36*	0.00	41.56	*80.73*	70.45	70.70
Treex								
CzEng 1.0	70.64	65.93	73.52	28.48	55.34	76.02	67.93	68.13
Treex CR	*75.99*	*81.63*	74.11	*45.37*	**60.87**	79.65	66.64	66.96

trained and tested on PDT 3.0 (Bejček et al., 2013)

	Relative	Reflexive	SzPP3	All
Count	1,075	579	1,950	3,604
Treex				
CzEng 1.0	57.14	67.57	50.52	55.20
Treex CR	78.40	76.19	61.31	**68.46**

MEASURE

- for an anaphor candidate a_i increment the counts:
  - *true*(a_i) if a_i is anaphoric in the gold annotation;
  - *pred*(a_i) if the CR system claims a_i is anaphoric;
  - *both*(a_i) if both the system and gold annotation claim a_i is anaphoric and the antecedent found by the system belongs to the transitive closure of all mentions coreferential with a_i in the gold annotation.

$$P = \sum_{a_i} \frac{both(a_i)}{pred(a_i)} \quad R = \sum_{a_i} \frac{both(a_i)}{true(a_i)} \quad F = \frac{2PR}{P+R}$$

CONCLUSION

- CR resolver for both Czech and English - addressing also zeros
- in Czech, it outperforms the original resolver used for CzEng 1.0
- in English, outperformed by state-of-the-art Stanford systems; however, good enough to be used for cross-lingual experiments

This project has been funded by the GAUK grant 338915 and the Czech Science Foundation grant GA-16-05394S.This work has been also supported and has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project No. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.