Incorporating Coreference to Automatic Evaluation of Coherence in Essays

Michal Novák, Kateřina Rysová, Magdaléna Rysová, Jiří Mírovský

Institute of Formal and Applied Linguistics
Charles University, Prague

Outline

1. Introduction
2. The original EVALD
3. Coreference-related extension to EVALD
4. Datasets
5. Experiments
6. Conclusion

Introduction

Task

- score the level of cohesion/coherence in essays
  - L1: essays written by native speakers
    - grades used in Czech elementary and high schools
    - highest → 1 2 3 4 5 ← lowest
  - L2: essays written by foreign learners of the language
    - proficiency levels as specified by The Common European Framework of Reference for Languages (CEFR)
    - lowest → A1 A2 B1 B2 C1 C2 ← highest
- motivation
  - a part of the system that can:
    - assist teachers with evaluation of essays
    - help learners to reveal possible errors

Example

Dobry den.
Jsemenuje se QQQ.
Ja student v UJOP.
Ja vstavam rano až v sedum hodin.
Jdu do školy tam studuju češtinu a angličtinu.
Je Dnes byla matematika.
Ja chodím do XXX.
Dalše jdu do domu.
Delam domašny ukolí.
Čtu knihý.
Pišu česky v sešite.
Ja večeřím maso a polevku.
Dalše jsu spát.
Jak každy XXX.

A1

- no use of discourse connectives at all
- simple sentences “I do something”
- the pronoun “I” is not dropped
- …
- quickly jumping from one topic to another with no details

Related work

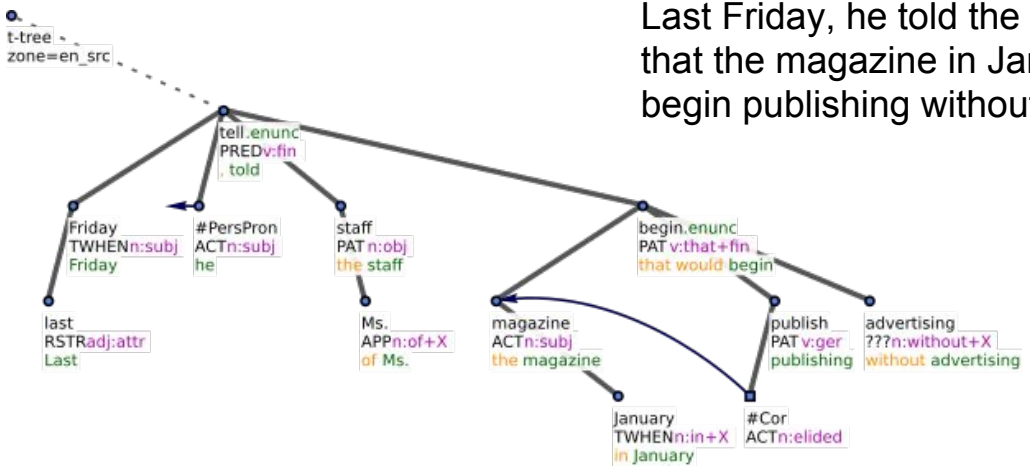
- Automated Essay Scoring
  - very old task: (Page, 1968)
  - using discourse and stylistic features: e-rater® (Attali and Burstein, 2006)
  - using coreference resolution: (Zupanc and Bosnić, 2017) and (Wonowidjojo et al., 2016)
- Proficiency Level Classification
  - Estonian (Vajjala and Lõo, 2014): F-score = 78.5%
  - Swedish (Pilán et al., 2016): F-score = 72%

The score is holistic, i.e. all aspects of the language are evaluated by a single common grade.

The original EVALD

EVALD

- EVALuator of Discourse 1.0 (Rysová et al., 2016)
- a system based on traditional ML
- using features describing frequency and variety of lexical and discourse items
- operates on automatically analysed texts
  - up to the level of deep syntax



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

Preprocessing

- Treex NLP Framework (Popel and Žabokrtský, 2010)
- Sentence splitting and tokenization
  - rule-based
- Part-of-speech tagging and morphology
  - MorphoDiTa tool (Straková et al., 2014)
- Dependency parsing
  - MST parser adapted to Czech (Novák and Žabokrtský, 2007)
- Surface-to-deep syntax transformation
  - Mainly rule-based
- Discourse parsing
  - Focused on local relations marked by explicit connectives
    - intra- and inter-sentential
  - Rule-based + exploiting lists of connectives and their discourse senses from Prague Discourse Treebank 2.0 (PDiT; Rysová et al., 2016)

Original features

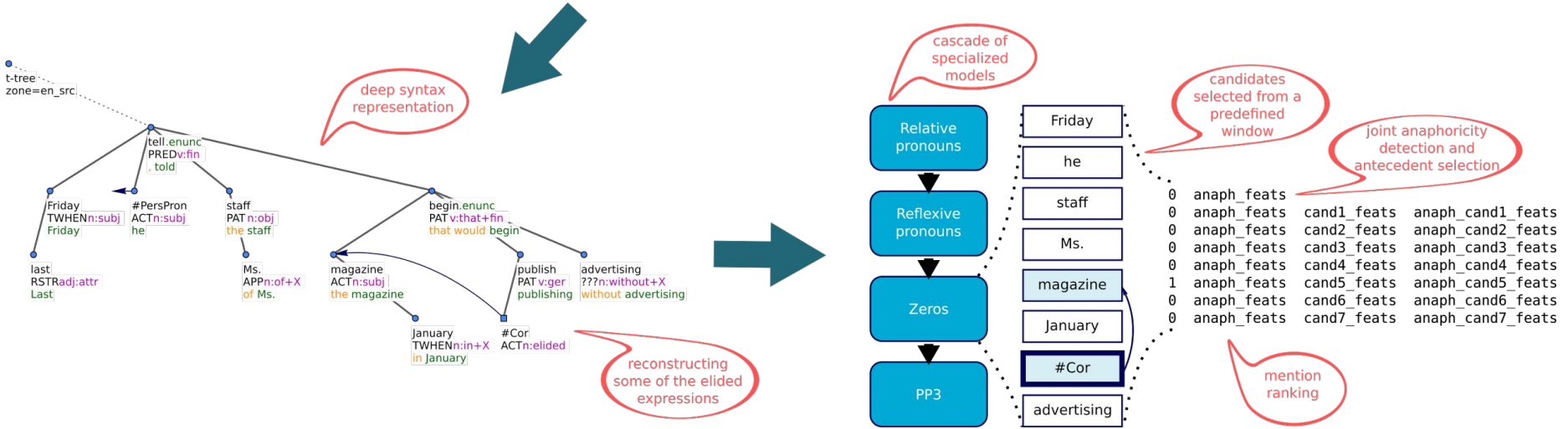
- Surface
  - Extracted from the tokenized text only
  - Lexical: number of tokens per sentence, Yule’s and Simpon’s index of lemmas diversity
  - Discourse: number of occurrences of any of 49 most frequent connectives in PDiT 2.0
- Advanced
  - Extracted from linguistically preprocessed data
  - Syntax: frequency of predicate-less sentences
  - Discourse: frequency of intra- and inter-sentential relations, the proportion of selected connectives in discourse relations, the proportion of 4 major types (temporal, contingency, contrast, expansion) in discourse relations

Coreference-related Extension to EVALD

Coreference Resolver

- Treex CR (Novák, 2017)
- F-score of finding any of the pronoun's antecedents: 68%

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



Coreference-related features

- Coreference features
  - Take advantage of the Treex CR
  - Quantitative
    - number of chains / links relative to the text length
    - distribution of chains by their length
    - proportion of intra- and inter-sentential links
  - Qualitative: variety of expressions forming the coreferential chains
    - lemmas
    - types of expressions (noun, zero, pronoun subtype)

Coreference-related features

- Pronoun features
  - No use of CR
  - They capture both anaphoric and non-anaphoric occurrences
  - Quantitative: relative frequency of pronouns and their subtypes
    - among all words / nouns and pronouns / pronouns
    - including zero subjects
  - Qualitative: how wide is the repertoire of used pronouns?
    - pronouns and zeros at the subject position
    - excessive use of a demonstrative pronoun “to” (“it/this/that”)

Datasets

Data sources

- Merlin (Boyd et al., 2014)
  - texts written by non-native speakers at CEFR exams
  - 441 texts
  - rated also with cohesion/coherence level
- CzeSL-SGT (Šebesta et al., 2014)
  - texts written by non-native speakers in courses of Czech for foreigners
  - 8,617 texts
  - no cohesion/coherence grades
- Skript2012 (Šebesta et al., 2016)
  - texts written by native speakers of Czech during the lessons of Czech language at elementary and high schools
  - 1,694 texts
  - no cohesion/coherence grades

Datasets

L1 - native speakers of Czech

- grades 1-5 (highest-lowest)
- formed using texts from Skript2012
- we manually labeled them with grades for a coherence/cohesion level

L2 - learners of Czech as a foreign language

- levels A1-C2 (lowest-highest)
- its core constituted by texts from Merlin
- less populated levels (A1, A2, C1) complemented with texts from CzeSL-SGT
- C2 level supplied with L1 texts

L1 dataset	**1**	**2**	**3**	**4**	**5**	**Total**
# documents	484	149	121	239	125	1,118
# sentences	20,986	4,449	2,913	3,382	939	32,669
# tokens	301,238	65,684	40,054	43,797	11,379	462,152

L2 dataset	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**Total**
# documents	174	176	171	157	105	162	945
# sentences	1,802	2,179	2,930	2,302	1,498	10,870	21,581
# tokens	15,555	21,750	27,223	37,717	21,959	143,845	268,049

Experiments

Evaluation Measures

- distribution of grades in the train/test data is artificial and we do not know the real distribution
- rather assume test data coming from a uniform distribution

- even a human judge has sometimes difficulties to determine the grade precisely

1. exact accuracy on balanced testset
    - test data balanced by sampling each class to the size of the smallest one
2. macro-averaged F-score
    - use all test data
    - calculate F-score for each class and average

$$F = \frac{1}{|C|} \sum_{c \in C} F_c$$

3. one-level tolerance accuracy on balanced testset
    - correct if a predicted grade is equal or neighboring to the true one

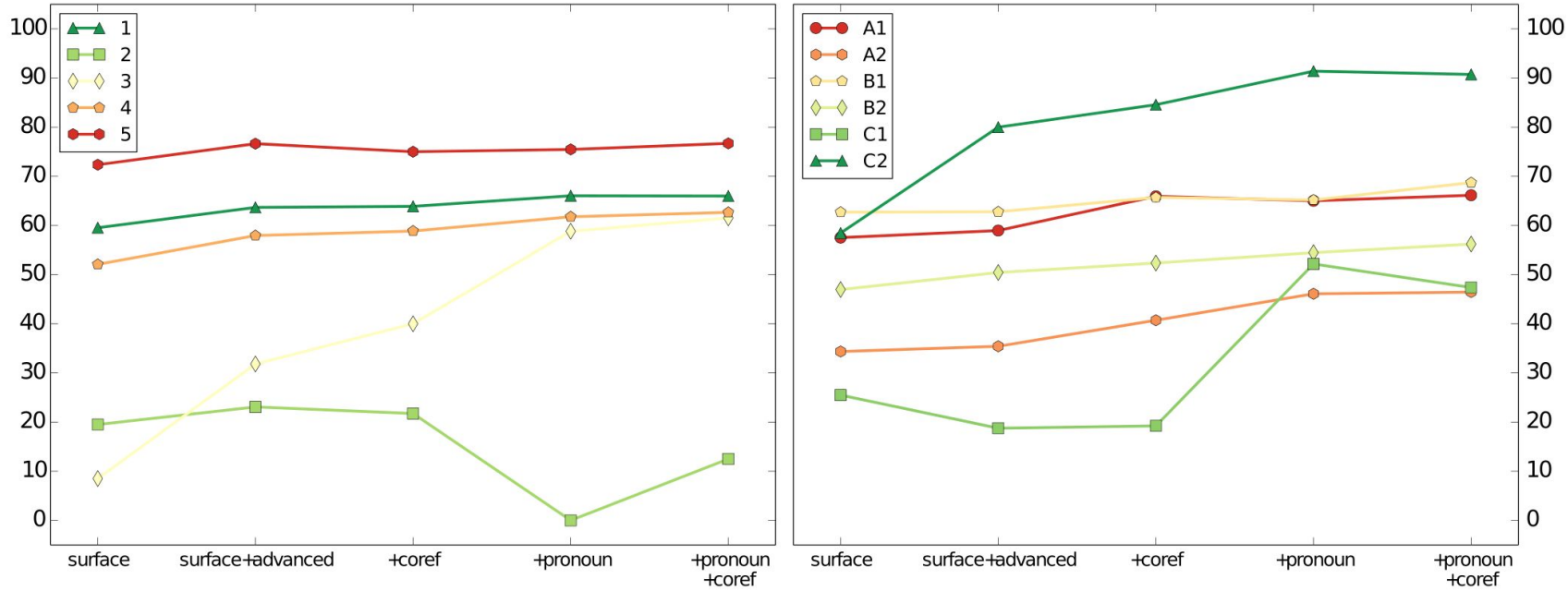
Experimental setup

- 10-fold cross-validation
- L1: 1,118 docs L2: 945 docs
- random forests
- baselines:
  - surface
  - surface + advanced
- system variants:
  - + pronoun
  - + coref
  - + pronoun + coref

Results

	L1 dataset			L2 dataset		
	F	e-Acc	1-Acc	F	e-Acc	1-Acc
surface	40.1	42.1	72.4	47.6	48.5	74.7
surface+advanced	44.9	46.1	80.8	51.3	55.5	82.5
+pronoun	45.9	48.2	**83.0**	58.6	62.3	**86.8**
+coref	45.2	47.0	81.3	54.7	58.7	85.2
+pronoun+coref	**46.0**	**49.5**	**83.0**	**59.0**	**63.3**	85.5

Analysis of the performance



Conclusion

Conclusion

- scoring level of cohesion/coherence in L1/L2 essays
- the coreference extension outperformed the original system by 3 and 5 percentage points for L1 and L2, respectively
- we collected two datasets that can be used for further experiments
- Future work:
  - more information in the output
  - include topic-focus articulation features

Thank You! Questions?