Automatic Classification of Russian Learner Errors

Alla Rozovskaya

Queens College, CUNY

Standard Reference-Based Evaluation for Grammatical Error Correction (GEC)

Source	The settings are very **reallistic** and the actors **had a great** performance .
Reference Gold (RG)	The settings were very realistic and the actors *gave* a great performance .
Hypothesis	The settings are very *realistic* and the actors *had great* performance.

Gold edits: (1) reallistic -> realistic;
(2) had -> gave
(3) are -> were

System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Correct edits: (1) reallistic -> realistic

Precision: 1/2=0.5
Recall: 1/3=0.33

No information on how the system performs on specific error types!

Classifying the edits

- For the gold edits, the annotators can be asked to provide a linguistic category during annotation

Gold edits:
(1) reallistic -> realistic ⇨ Spelling
(2) a -> Ø ⇨ Determiner
(3) are -> were ⇨ Verb tense

Classifying system edits

- State-of-the-art GEC systems are based on neural machine translation (NMT) architecture
- Classifying system edits is not trivial since the systems are not restricted in the types of edits that can be made

3

The problem

- System performance can only be evaluated overall on all edits
- Automatic edit classification is necessary to perform type-based evaluation of system performance
- Type-based evaluation
  - Can provide insight into further system development
  - Is necessary in order to provide useful feedback to language learners, when a mistake is identified
  - Allows for a standardization of multiple GEC datasets that may have been annotated with different error taxonomies

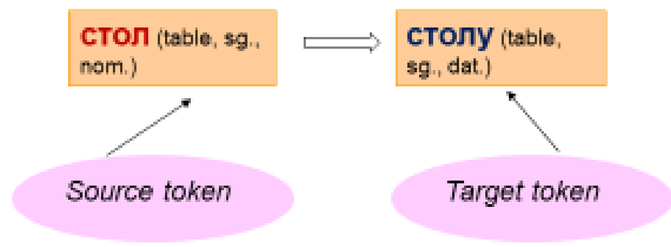
Error classification tool for Russian

- Our approach is inspired by ERRANT
  - We use POS and morphological information to classify edits
- Adapted to the specific challenges of Russian
- The tool is applied to classify edits in two Russian learner corpora
  - Manual evaluation with human annotators reveals that the accuracy of the edit classification is 93%
- The tool is applied to 2 GEC systems
  - Type-based performance evaluation shows "easy" and "challenging" errors in Russian GEC

Overview of the rules

- Morphological analyzer is applied to source and target token of an edit

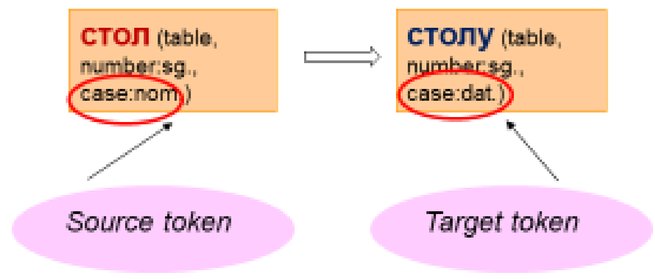
Sample edit:



The analyzer produces:
- base form
- list of morphological properties (e.g., number, gender, case, aspect, voice, person, tense)

Predicting error type

- **When base form is the same** on source and target, error type is predicted based on the morphological property that has different values in the source and target token



- **When base forms are different**
  - Predict spelling error if the source word is not in the dictionary
  - Predict lexical error otherwise
- See paper for more details on the rules.

Challenges specific to Russian

- Some Russian surface forms have **multiple analyses**
  - E.g. sg., gen. can be confused with pl., nom.
- Such cases are problematic since depending on the chosen analysis a different mismatch in the grammatical category (case or number) will be identified
- We handle these cases by predicting 2 mismatch categories

Manual evaluation of automatic error categories

Rater	RULEC			RU-Lang8		
	Good	Accept.	Bad	Good	Accept.	Bad
1	70	25	5	88	8	4
2	63	25	12	83	12	5

Manual evaluation of the automatically assigned error categories by each rater and each dataset on a set of 100 edits, randomly selected.

Type-based evaluation

Error type	CNN			Transformer		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Spelling	66.2	53.9	63.3	75.93	63.73	**73.13**
Lex. choice	46.3	3.0	12.1	67.07	13.43	**37.29**
Punc.	54.8	23.3	**43.1**	42.71	6.93	21.01
Replace	0.00	0.00	0.00	2.30	1.05	1.86
Prep.	25.2	8.1	17.7	70.25	25.53	**52.02**
Morph.	20.0	1.8	6.7	51.61	14.55	**34.19**
Insert	0.0	0.0	0.0	17.39	6.35	**12.90**
Delete	75.0	3.2	13.6	38.24	13.83	**28.26**
Noun (all)	61.1	38.4	54.6	72.0	36.4	**60.2**
Verb (all)	54.5	20.4	40.8	71.5	38.0	**60.8**
Adj (all)	50.0	21.6	39.6	64.1	29.5	**51.9**

Type-based evaluation on the RULEC dataset.