How Far are We from Fully Automatic High Quality Grammatical Error Correction?

od autorů Christophera Bryanta a Hwee Tou Nga

Markéta Masopustová

12. 5. 2016

Úvod

- nárůst zájmu o tematiku gramatických korektorů chyb (GEC), ale chybí robustní evaluační metoda
- výsledky systémů se porovnávají s ruční anotací jednoho anotárora *gold standard annotations* (GSA)
- v rámci experimentu je porovnávána shoda více anotátorů v GSA s výstupem z GEC systému
- návrh nové metody evaluace

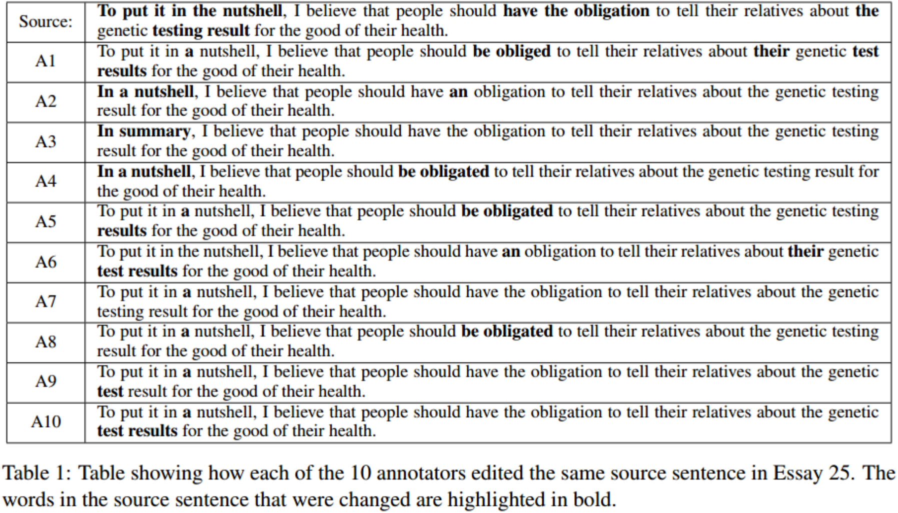
Mezianotátorská shoda

- data s větší shodou jsou důveryhodnější
- často ve využívá Cohenovo κ
- funguje dobře pro možnosti, které mají jednu správnou a dobře definovanou variantu
  - part-of-speech tagging

Mezianotátorská shoda v GEC

- příklad výzkumu Tetrault a Chodorow (2008)
  - dva rodilí mluvčí
  - 200 vět, ve kterých chyběla pouze jedna předložka
  - κ 0,7
  - vyjadřuje nízká K opravdu úplnou neshodu?
- jiný výzkum Rozovskaya and Roth (2010)
  - tři anotátoři
  - 200 vět, je správná či nikoli (YES/NO)
  - κ 0,16, 0,4 a 0,23
  - obtížnost úkolu a rozdílné názory na správnost
- podobný problém s IAA ve statistickém strojovém překladu

Ukázka



Source:	**To put it in the nutshell**, I believe that people should **have the obligation** to tell their relatives about **the** genetic **testing result** for the good of their health.
A1	To put it in **a** nutshell, I believe that people should **be obliged** to tell their relatives about **their** genetic **test results** for the good of their health.
A2	**In a nutshell**, I believe that people should have **an** obligation to tell their relatives about the genetic testing result for the good of their health.
A3	**In summary**, I believe that people should have the obligation to tell their relatives about the genetic testing result for the good of their health.
A4	**In a nutshell**, I believe that people should **be obligated** to tell their relatives about the genetic testing result for the good of their health.
A5	To put it in **a** nutshell, I believe that people should **be obligated** to tell their relatives about the genetic testing **results** for the good of their health.
A6	To put it in the nutshell, I believe that people should have **an** obligation to tell their relatives about **their** genetic **test results** for the good of their health.
A7	To put it in **a** nutshell, I believe that people should have the obligation to tell their relatives about the genetic testing result for the good of their health.
A8	To put it in **a** nutshell, I believe that people should **be obligated** to tell their relatives about the genetic testing result for the good of their health.
A9	To put it in **a** nutshell, I believe that people should have the obligation to tell their relatives about the genetic **test** result for the good of their health.
A10	To put it in **a** nutshell, I believe that people should have the obligation to tell their relatives about the genetic **test results** for the good of their health.

Table 1: Table showing how each of the 10 annotators edited the same source sentence in Essay 25. The words in the source sentence that were changed are highlighted in bold.

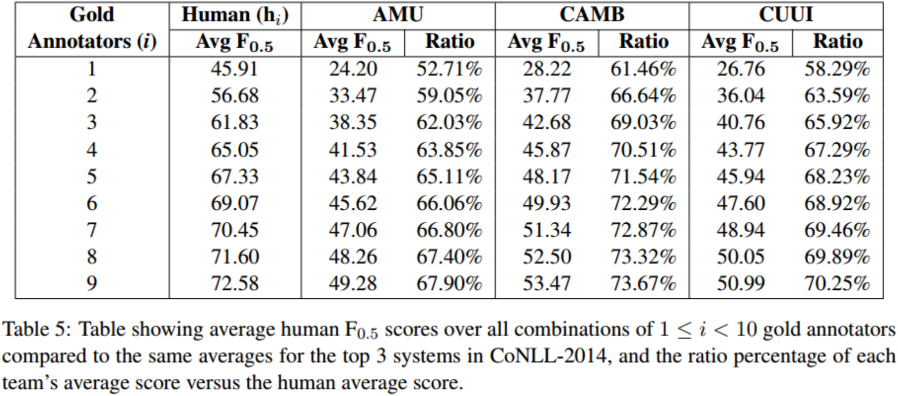
Kolekce dat

- 25 studentů z National University of Singapore – nejsou rodilí mluvčí angličtiny
- celkem 50 esejí se zhruba stejnou délkou a kvalitou
- 10 anotátorů – učitelé angličtiny, korektoři, lingvisti
- označení chyby, její oprava a kategorizace

Kvantitativní analýza

- analýza dat a jejich porovnání s daty z *Conference on Natural Language Learning 2014* pomocí *Max-Match Scorer* (Dahlmeier and Ng, 2012)
- systém, který vyhodnocuje na úrovni vět v případě, že jde o opravy, návrhy úprav a *gold* úpravy a počítá $F_{0,5}$ míru
  - narazí-li na více *gold* pravidel, spočítá $F_{0,5}$ míru pro všechny a vybere tu nejvyšší
  - $F_{0,5}$ míra se počítá proto, že chceme v ideálním případě opravit všechny chyby – přesnost (precision) je tedy důležitější než pokrytí (recall)

Vyhodnocení



Gold	Human (h_i)	AMU		CAMB		CUUI	
Annotators (i)	Avg $F_{0.5}$	Avg $F_{0.5}$	Ratio	Avg $F_{0.5}$	Ratio	Avg $F_{0.5}$	Ratio
1	45.91	24.20	52.71%	28.22	61.46%	26.76	58.29%
2	56.68	33.47	59.05%	37.77	66.64%	36.04	63.59%
3	61.83	38.35	62.03%	42.68	69.03%	40.76	65.92%
4	65.05	41.53	63.85%	45.87	70.51%	43.77	67.29%
5	67.33	43.84	65.11%	48.17	71.54%	45.94	68.23%
6	69.07	45.62	66.06%	49.93	72.29%	47.60	68.92%
7	70.45	47.06	66.80%	51.34	72.87%	48.94	69.46%
8	71.60	48.26	67.40%	52.50	73.32%	50.05	69.89%
9	72.58	49.28	67.90%	53.47	73.67%	50.99	70.25%

Table 5: Table showing average human $F_{0.5}$ scores over all combinations of $1 \leq i < 10$ gold annotators compared to the same averages for the top 3 systems in CoNLL-2014, and the ratio percentage of each team's average score versus the human average score.

Shrnutí

- ani člověk v porovnání s člověkem nesáhne shody 100 %, většinou kolem 73 % $F_{0,5}$ míry
- nejlepší týmy kolem 67–73 % $F_{0,5}$ míry

Zdroje

- Christopher Bryant and Hwee Tou Ng. 2015. How Far are We from Fully Automatic High Quality Grammatical Error Correction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing: Association for Computational Linguistics, s. 697–707.
- Joel R. Tetrault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*, pages 24–32, Manchester, UK.
- Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.
- Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *HLTNAACL*, pages 568–572.