EBERHARD KARLS UNIVERSITÄT TÜBINGEN

SFB 833
BEDEUTUNGSKONSTITUTION
TRANSFERPROJEKT T1

The Impact of Spelling Correction and Task Context on Short Answer Assessment for Intelligent Tutoring Systems

Ramon Ziai, Florian Nuxoll, Kordula De Kuthy, Björn Rudzewitz & Detmar Meurers

8th NLP4CALL Workshop, Turku

September 30, 2019

Introduction

- Short Answer Assessment (SAA): the task of determining whether a given response to a question is acceptable or not
  - Often called Automatic Short Answer Grading (ASAG) in cases where the outcome is on ordered scale (numeric score)
- Field has attracted considerable attention:
  - Shared tasks: ASAP-SAS 2012 on Kaggle, Task 7 at SemEval 2013
  - Recent approaches: Riordan et al. (2017); Gomaa and Fahmy (2019)
- SAA can however not be considered a solved problem:
  - Still unclear how well standard SAA approaches work in real-life educational contexts, such as

  → integrating language tutoring systems into a regular school setting.

Motivation

- In tutoring systems, the goal is to give immediate feedback on the language produced by the learner
  - e.g. help students complete homework exercises in the system step by step.
- Especially challenging for comprehension exercises:
  - System needs to evaluate the meaning provided by the student response, and possibly give helpful feedback for improvement
- SAA can help with the evaluation part:
  - If an answer is deemed correct, the feedback is positive,
  - if not, further diagnosis can be carried out.

Goals

- We report on SAA work in progress on authentic data from a language tutoring system for 7th grade English.
- We employ an alignment-based SAA system (CoMiC, Meurers, Ziai, Ott, and Bailey 2011a)
  - Shown to work well for several data sets where target answers are available (Meurers et al. 2011b; Ott et al. 2013)
- We investigate two main factors for SAA performance:
  1. The impact of automatic **spelling normalization** on SAA using a noisy channel approach (Brill and Moore 2000)
  2. The influence of **different test scenarios**, namely ‘unseen answers’, ‘unseen items’, and ‘unseen tasks’ (cf. Dzikovska et al. 2013)

Data

- Our data comes from the FeedBook
  (Rudzewitz et al. 2017, 2018; Ziai et al. 2018)
  - English tutoring system for 7th grade used in German secondary schools as part of a full-year randomized controlled field study (Meurers et al. 2019)
- The system includes interactive feedback on form for all grammar topics on the curriculum,
  - and also a first version of meaning feedback for meaning-oriented tasks, such as reading and listening comprehension activities.
- This enabled the collection of data from student-system interactions on comprehension tasks.

Example

1 On the move

B 7 Talking to Gwynn

b) Listen again and complete the statements in 1 to 3 words.

0:00 / 2:11

1. Gwynn tells Mrs Collins that Gillian needs *time* ✔ to get used to the situation.
2. Mrs Collins thinks Gillian should try to be ______ towards Gwynn.
3. Gwynn thinks Gillian feels desperate because she doesn't want to ______ .
4. Gwynn suggests that Mrs Collins should ______ on her own.
5. Gwynn thinks Gillian is most worried about ______ when she moves to Wales.
6. Gwynn suggests that Gillian can come to Wales for a weekend and invite ______ .

Resulting Data Set

- We extracted all student responses that were entered in reading or listening comprehension tasks, filtering out
  - duplicate answers,
  - answers to tasks that were erroneously classified as meaning-oriented or
  - that require knowledge external to the task material.
- Result: 3,829 answers entered into 123 answer fields of 25 tasks, on average 7.11 tokens long
  - Distribution uneven, almost 40% of the answers from one task
  - The nine gap-filling tasks typically triggered shorter responses than the 16 tasks with sentential input
- An experienced English teacher rated every response with respect to whether it is an acceptable answer or not.

Spelling Correction

- Our spelling correction approach is based on the noisy channel model described by Brill and Moore (2000)
  - Implementation by Adriane Boyd: `https://github.com/adrianeboyd/BrillMooreSpellChecker`
- Requirements:
  - A list of misspellings (non-word/correction pairs) to derive the model
  - A dictionary of valid words to use as corrections
- We trained the approach on a list of approximately 10,000 misspellings made by German learners of English
  - extracted from the EFCamDat corpus (Geertzen et al. 2013)

Task-aware Spelling Correction

- The dictionary was compiled from the vocabulary list of English school books used in German schools up to 7th grade
  - approximating the vocabulary that German 7th graders learning English in a foreign language learning setting were exposed to.
- Task-awareness is achieved by weighting dictionary entries:
  - Weight of 1 for standard entries
  - Increased by term frequency in the specific task's reading or listening text

→ Task-specific spelling corrections are more likely to happen, given a sufficiently close learner production.

Experiment Setup

- We employed a variant of the CoMiC system (Meurers, Ziai, Ott, and Bailey 2011a)
  - Aligns different linguistic units (tokens, chunks, dependencies) of the learner and the target answers to one another
  - Extracts numeric features based on the number and type of alignments found
  - Features are then used to train a classifier for new unseen answers
- We used a Support Vector Machine (SVM) with a polynomial kernel as the classification approach
  - based on the *kernlab* package (Karatzoglou et al. 2004) in *R* (R Core Team 2015) via the *caret* machine learning toolkit (Kuhn 2008)
  - We used default hyperparameters for the SVM approach.

Experiment Setup II

- Baseline system: nine standard string similarity measures from the *stringdist* package (van der Loo 2014) in *R*,
  - Similarity scores calculated between student and target response were used in the same classification setup as the CoMiC features.
- Spelling correction was incorporated as a pre-processing step
  - → second version of CoMiC enhanced with spelling correction
  - Apart from this pre-processing, the two CoMiC versions are identical
- We used the following test scenarios (cf. Dzikovska et al. 2013):
  - 'unseen answers': tenfold cross-validation across all answers
  - 'unseen items': for each item, all answers for that item (gap/field) are held out; training is done on all other answers.
  - 'unseen tasks': for each task, all answers for that task are held out

Overall Results

SAA System	Unseen answers %	Unseen answers κ	Unseen items %	Unseen items κ	Unseen tasks %	Unseen tasks κ
Majority	62.05%, κ = 0.00					
stringsim	78.35	0.52	76.97	0.48	75.61	0.45
CoMiC	81.25	0.59	81.20	0.59	80.80	0.58
+SC	**82.63**	**0.62**	**82.63**	**0.61**	**82.45**	**0.61**

- String similarity model surprisingly strong
  $\rightarrow$ many real-life cases can be scored with surface-based methods
- Majority baseline and string similarity model are clearly outperformed by CoMiC.
  - Higher level of linguistic abstraction allows for better generalization
- Spelling Correction (+SC) leads to systematic improvement
- ‘Unseen tasks’ most challenging, but also closest to real life

Unseen Tasks (top 10, sorted by # answers)

Task ID	input	type	# answers	∅ tokens	CoMiC		CoMiC+SC	
					%	κ	%	κ
2B1	gap-filling	reading	1,511	7.04	80.15	0.53	**82.46**	**0.57**
3A3a	sentence(s)	reading	463	9.77	79.70	0.53	**82.51**	**0.58**
1CYP2b	sentence(s)	listening	411	7.83	**88.32**	0.71	88.08	0.71
1ET5	sentence(s)	reading	360	4.68	93.33	0.86	**93.61**	**0.87**
2CYP3	sentence(s)	reading	255	7.71	72.94	0.45	**75.29**	**0.49**
1B7b	gap-filling	listening	220	1.79	64.09	0.29	**70.45**	**0.42**
2C5b	sentence(s)	reading	177	9.24	84.75	0.69	**85.88**	**0.72**
1AP37	sentence(s)	reading	126	8.90	**73.81**	**0.44**	70.63	0.38
1AP38	sentence(s)	reading	85	14.15	87.06	0.74	87.06	0.74
2ET3	gap-filling	reading	61	2.59	**62.30**	**0.25**	54.10	0.10

- Positive impact of spelling correction for most tasks, but not all
- What makes it work or not work?

Negative effects: Mal-corrections

- For some tasks, spelling correction mal-corrected answers into worse versions
- Example:

  (1) Prompt: ‘Robin ran away because of trouble with his father.’
  A_{orig}: ’Robin ran away because of trouble with his stepfather.’
  A_{corr}: ’Robin ran away because of trouble with his stepmother.’

- Cause: ‘stepfather’ apparently not in dictionary

→ Dictionary needs to be extended to include plausible alternatives to explicitly mentioned material

Positive effects: Hard-to-spell words

- We manually inspected some student responses for task ‘2B1’.
- Many spelling corrections revolved around Welsh proper names, such as ‘Gruffudd’ or ‘Llandysul’.
  → Very hard to spell for 7th grade English learners, but successfully corrected by our spelling correction approach
- Effect of spelling correction possibly connected to the lexical material involved in the task, instead of formal properties

→ Systematic analysis of lexical complexity and/or complex word identification in task texts could be promising
(see e.g. Yimam et al. 2018)

Conclusion

- We presented work in progress on Short Answer Assessment (SAA) on data from the FeedBook,
  - an English language tutoring system we employed in a real-life school setting in Germany.
- To investigate the influence of spelling correction on SAA, we added a noisy channel model to a standard SAA approach
  - Result: general increase of classification performance for the data we collected
- A Task-by-task analysis revealed that the effect of spelling correction is not uniform across tasks.
  - May be related to lexical characteristics of the language employed in the task context
  - Systematical analysis of lexical complexity and integration of complex word identification could verify this hypothesis.

References

Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Hong Kong, October 2000. ACL. doi: 10.3115/1075218.1075255. URL https://www.aclweb.org/anthology/P00-1037.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S13-2045.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Cascadilla Press, 2013. URL http://purl.org/icall/efcamdat.

Wael Hassan Gomaa and Aly Aly Fahmy. Ans2vec: A scoring system for short answers. In Aboul Ella Hassanien, Ahmad Taher Azar, Tarek Gaber, Roheet Bhatnagar, and Mohamed F. Tolba, editors, *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)*, pages 586–595, Cham, 2019. Springer International Publishing. ISBN 978-3-030-14118-9.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL http://www.jstatsoft.org/v11/i09/.

Max Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. ISSN 1548-7660. doi: 10.18637/jss.v028.i05. URL https://www.jstatsoft.org/v028/i05.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation*, 21(4):355–369, 2011a. URL http://purl.org/dm/papers/Meurers.Ziai.ea-11.pdf.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, 2011b. URL http://aclweb.org/anthology/W11-2401.pdf.

Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188, 2019. URL https://doi.org/10.1017/S0267190519000126.

Niels Ott, Ramon Ziai, Michael Hahn, and Detmar Meurers. CoMeT: Integrating different levels of linguistic modeling for meaning assessment. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval)*, pages 608–616, Atlanta, GA, 2013. ACL. URL http://aclweb.org/anthology/S13-2102.pdf.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, Copenhagen, Denmark, September 2017. ACL. doi: 10.18653/v1/W17-5017. URL https://www.aclweb.org/anthology/W17-5017.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition*, 2017. URL http://aclweb.org/anthology/W17-0305.pdf.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. Generating feedback for English foreign language exercises. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 127–136. ACL, 2018. URL http://aclweb.org/anthology/W18-0513.pdf.

M.P.J. van der Loo. The stringdist package for approximate string matching. *The R Journal*, 6: 111–122, 2014. URL https://CRAN.R-project.org/package=stringdist.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June 2018. ACL. doi: 10.18653/v1/W18-0507. URL https://www.aclweb.org/anthology/W18-0507.

Ramon Ziai, Björn Rudzewitz, Kordula De Kuthy, Florian Nuxoll, and Detmar Meurers. Feedback strategies for form and meaning in a real-life language tutoring system. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted*

Language Learning (NLP4CALL), pages 91–98. ACL, 2018. URL
`http://aclweb.org/anthology/W18-7110`.