Automated Text Classification to Capture Scientific Reasoning and Argumentation Processes in Different Professional Problem Solving Contexts

Andras Csanadi[1], Johannes Daxenberger[2], Christian Ghanem[1], Ingo Kollar[1,3], Frank Fischer[1], Iryna Gurevych[2]
[1] Munich Center of the Learning Sciences, LMU Munich, reason@psy.lmu.de
[2] Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt, http://www.ukp.tu-darmstadt.de
[3] University of Augsburg, ingo.kollar@phil.uni-augsburg.de

Abstract

This paper explores the possibilities to automatically code scientific reasoning and argumentation (SRA). Our empirical work extends two previous studies. Those studies used transcribed verbal data to develop a reliable coding scheme on SRA in the domains of teacher education and social work. In the present paper we introduce the results of an automated coding system we developed to assess SRA in these two domains. We discuss within- and cross-domain experiments and consider further improvements.

Keywords: scientific reasoning and argumentation, automated text classification, cross-domain generalizability

Background

Scientific reasoning and argumentation (SRA) is a complex process (Kuhn et al., 2000) which makes its assessment challenging for instructional design. Moreover, with the growth of computer-based and online learning platforms (in particular, MOOCs), the need to automatically assess SRA processes has become more widespread among course-developers so as to provide adaptive support (Dyke et al., 2013) for learners. The present empirical work investigates the capabilities and limitations of automated text classification for SRA processes across different domains of practice. The SRA framework (Fischer et al., 2014) we applied identifies eight epistemic activities of SRA. While solving problems, reasoners might (1) identify the problem itself; (2) develop questions for further investigation; (3) set hypotheses; (4) generate evidence; (5) evaluate evidence; (6) draw conclusions; (7) create artefacts; and (8) communicate their ideas with others. Two previous studies developed a coding scheme to capture epistemic activities of SRA in the domains of teacher education (Csanadi, Kollar & Fischer, 2015) and social work (Ghanem et al., 2015). Their findings suggest that the epistemic activities of SRA can be reliably captured during practitioners' problem solving. Recent developments have given rise to computer-supported scientific reasoning assessments (e.g., Gobert et al., 2013). Automated text classification techniques developed by computer scientists in collaboration with domain experts may facilitate the process of text and discourse analysis (Kelly et al., 2015; Mayfield & Rosé, 2013). The novelty of the present research is the development of an automated text classification system that is based on fine-grained semantic information (e.g., discourse markers) to capture epistemic activities of scientific inquiry. This paper investigates (1) the potential of automated classification of SRA in professional problem solving; and (2) whether SRA can be reliably assessed across different domains.

Method

The present study relies on two German datasets from earlier studies. Study 1 (Csanadi et al., 2015) represents an experimental study with 39 teacher students discussing a problem case from their future practice for ten minutes, either alone or in pairs. In Study 2 (Ghanem et al., 2015), 22 social work students and 26 probation officers were asked to think aloud individually about a problem case (5-10 minutes on average). Both studies followed the coding recommendations of Chi (1997) and

Strijbos et al. (2006). First, audio data were transcribed. Then, a segmentation procedure divided the text into propositional units (80 - 85% reliability). Finally, a coding scheme was utilized that the authors developed to capture the eight epistemic activities of SRA (Fischer et al., 2014). Their coding approach resulted in acceptable reliability: κ = 0.68 (Study 1) and κ = 0.69 (Study 2). After segmentation, Study 1 contained approximately 2,700 units, and Study 2 contained approximately 4,500 units.

For the automatic coding process, we used DKPro TC[1] with Conditional Random Fields (CRF; Okazaki, 2007) to model the sequential nature of the SRA process. The CRF model is trained on a set of lexical (e.g., word n-grams) and dictionary features (LIWC classes for German; Wolf 2008), syntactic features (e.g., part-of-speech tags), discourse features (e.g., discourse markers; Eckle-Kohler et al., 2015), and semantic features (e.g., semantic domain labels from GermaNet). We carried out both within-domain experiments (10-fold cross-validation on the entire data from Study 1 resp. Study 2) as well as cross-domain experiments (training on entire data from Study 1 and testing on the entire data from Study 2 and vice versa). For each unit, the CRF had to predict one of nine codes (8 epistemic, 1 non-epistemic). We tested different types of features individually, and added a baseline predictor which always predicts the most frequent code from the training data.

Results

The results are displayed in Table 1. For both studies, the within-domain experiments clearly outperformed the baseline. Study 1 yielded slightly better scores when micro-averaging over units (accuracy), while for Study 2 the macro-averaged scores over the nine epistemic activities (F1) are better. Cross-domain experiments yielded drastically lower performance, in the case of training on Study 2 data and testing on Study 1 data, the results could not outperform the baseline. From the feature types we considered, lexical and syntactic features work best, followed by discourse features. Semantic features did not help too much. A combination of all features yielded the best results on both datasets (within-domain).

[1] https://github.com/dkpro/dkpro-tc

Table 1

Accuracy (upper) and macro-averaged F1-scores (lower) for each experiment.

		Feature types, CRF					
	Dataset	All	Lexical	Syntactic	Discourse	Semantic	Baseline
Within-domain	Teacher students	0.52	0.48	0.49	0.40	0.35	0.36
		0.34	0.31	0.27	0.15	0.06	0.06
	Probation officers	0.46	0.45	0.42	0.34	0.27	0.30
		0.39	0.37	0.34	0.16	0.12	0.05
Cross-domain	Teacher students to probation officers	0.18	0.16	0.20	0.03	0	0
		0.12	0.11	0.11	0.03	0	0
	Probation officers to teacher students	0.20	0.19	0.23	0.20	0.21	0.23
		0.13	0.13	0.13	0.11	0.07	0.04

Discussion

Our research intended to examine (1) whether an automated classification system can be applied to capture SRA processes (2) across two practical domains. The automated coding system trained on verbal data stemming from two earlier studies resulted in relatively low accuracy rates when trained and tested on data from the same domain. One reason was the rather small number of instances for the training phase and the very uneven distribution of epistemic activities. Moreover, the text originated from transcribed speech where informal grammar leads to noise for some of the linguistic features we considered. Finally, the automated coding did not work equally well for each of the activities included in the coding scheme; resulting in substantial confusion between some of the epistemic classes. The cross-domain experiments yielded lower reliability scores. This can be explained with the overall importance of lexical features (particularly for the data in Study 2) for the predictor. The performance of lexical features drops substantially in the cross-domain setup, which is to be expected due to the change of domain. The results can indicate at least moderate applicability of our automated classification system. However, future work should further clarify this question by including more training data; and more non-lexicalized features that are less prone to grammatical noise. Additionally, investigating verbal data from further domains could tell more about the cross-domain generalizability of our findings.

Acknowledgements

This research was supported by the Elite Network of Bavaria under Grant K-GS-2012-209, by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1416B (CEDIFOR) and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant № I/82806.

References

Chi, M. (1997). Quantifying Qualitative Analyses Of Verbal Data: A Practical Guide. *Journal of the Learning Sciences*, *6*(3), 271-315.

Csanadi, A., Kollar, I., & Fischer, F. (2015, August) Internal scripts and social context as antecedents of teacher students' scientific reasoning. Paper presented at the 16th Biennial Conference of the European Association for Research on Learning and Instruction (EARLI), Limassol, Cyprus.

Dyke, G., Adamson, D., Howley, I., & Rosé, C. P. (2013). Enhancing scientific reasoning and discussion with conversational agents. *Learning Technologies, IEEE Transactions on*, *6*(3), 240-247.

Eckle-Kohler, J., Kluge, R., & Gurevych, I. (2015, September). On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2249-2255, Lisbon, Portugal.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R.,. . . Eberle, J. (2014). Scientific Reasoning and Argumentation: Advancing an Interdisciplinary Research Agenda in Education. Frontline Learning Research, 5, 28–45.

Ghanem, C., Pankofer, S., Fischer, F., Kollar, I. & Lawson, T. R. (2015, April). The Relation between Social Work Practice and Science - Analysis of Scientific Reasoning of Probation Officers and Social Work Students. Paper presented at the European Conference for Social Work Research, Lubljana, Slovenia.

Kelly, N., Thompson, K., & Yeoman, P. (2015). Theory-led design of instruments and representations in learning analytics: Developing a novel tool for orchestration of online collaborative learning. *Journal of Learning Analytics*, *2*(2), 14-43.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*(4), 495-523.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite.

Mayfield, E., & Rosé, C. P. (2013). Open Source Machine Learning for Text. *Handbook of automated essay evaluation: Current applications and new directions*.

Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about?. *Computers & Education*, *46*(1), 29-48.

Wolf, M., Horn, A., Mehl, M., Haug, S., Pennebaker, J. W. & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, *53*(2), 85-98.