EBERHARD KARLS UNIVERSITÄT TÜBINGEN

SFB 833

Short Answer Assessment: Establishing Links Between Research Strands

Ramon Ziai, Niels Ott, Detmar Meurers
Collaborative Research Center SFB 833, Department of Linguistics, University of Tübingen

Motivation

- Short answer assessment systems have been developed for a range of purposes, on various data sources, employing different techniques.
- While clearly related, many approaches remain isolated.
- We sketch the landscape of short answer assessment, characterizing existing systems and their properties.
- In order to foster development and to connect research strands, more data sets and systems should be made available.
- Comparing two concrete systems on an available data set, we explore the issues involved in comparing such diverse systems in general.

Comparability of Approaches & Datasets

Datasets

- For results to be reproducible and to support serious system comparison, datasets must be publicly available. However, data sets also differ in
  - data source: reading comprehension task in language learning, tutoring system, automated grading of exams
  - language properties: native vs. learner language, domain-specific language (e.g., computer science)
  - assessment scheme: nominal vs. interval scale
- ➥ For meaningful comparison, data availability combined with explicit modeling of its source, properties, and classification scheme are crucial.

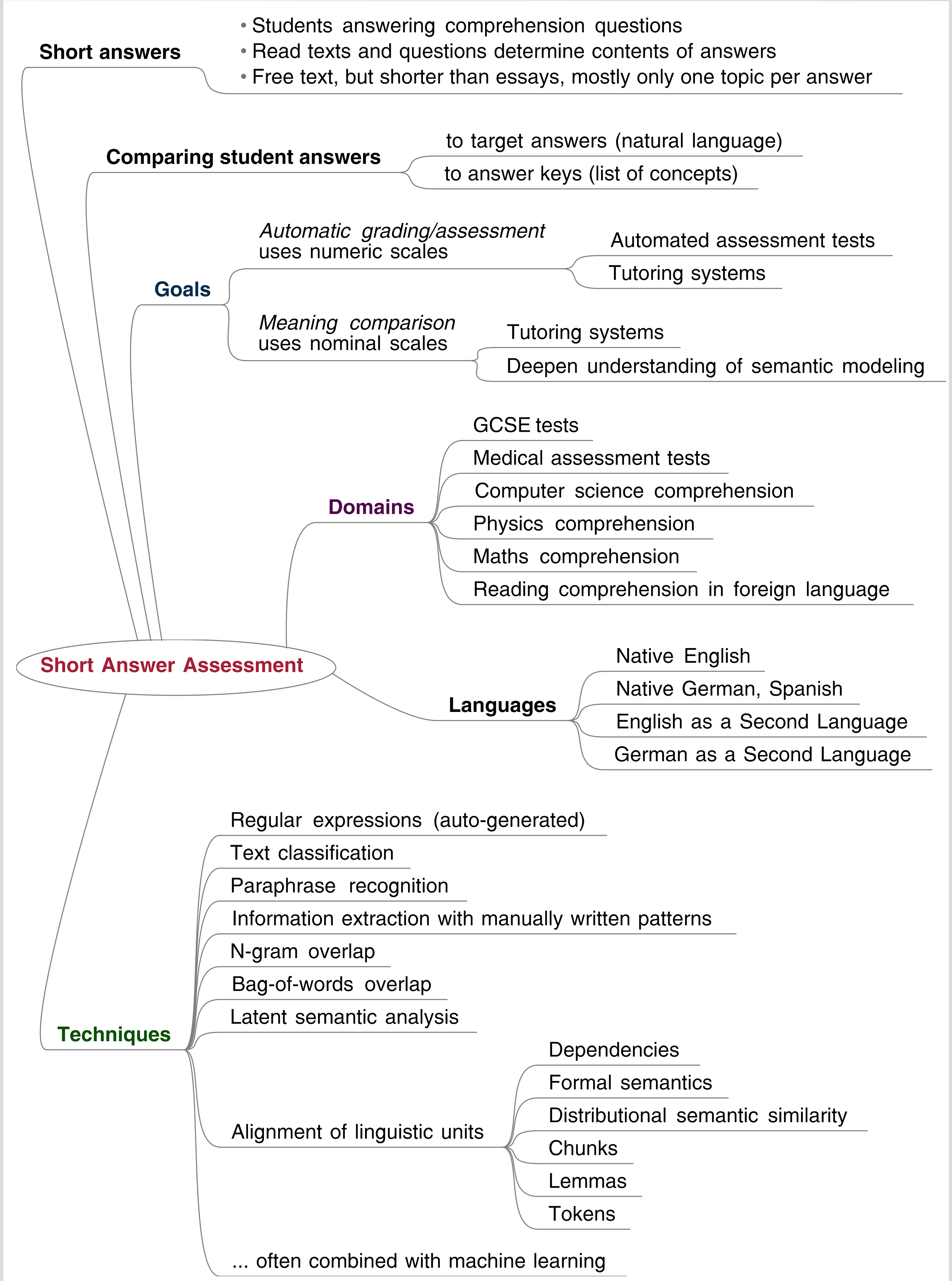
Evaluation Metrics

- Scoring systems are often evaluated using a pairwise correlation metric, whereas meaning comparison is associated with accuracy.
  - However, such correlation metrics assume a normal distribution and many datasets are biased towards correct answers.
  - Correlation generally suffers from low variance in gold ratings.
- Mohler et al. (2011) suggest RMSE as a remedy to capture a system's average error in scoring.
  - But RMSE is dependent on task and scale and thus does not support comparing studies differing in these aspects.
- ➥ Best to report multiple measures.

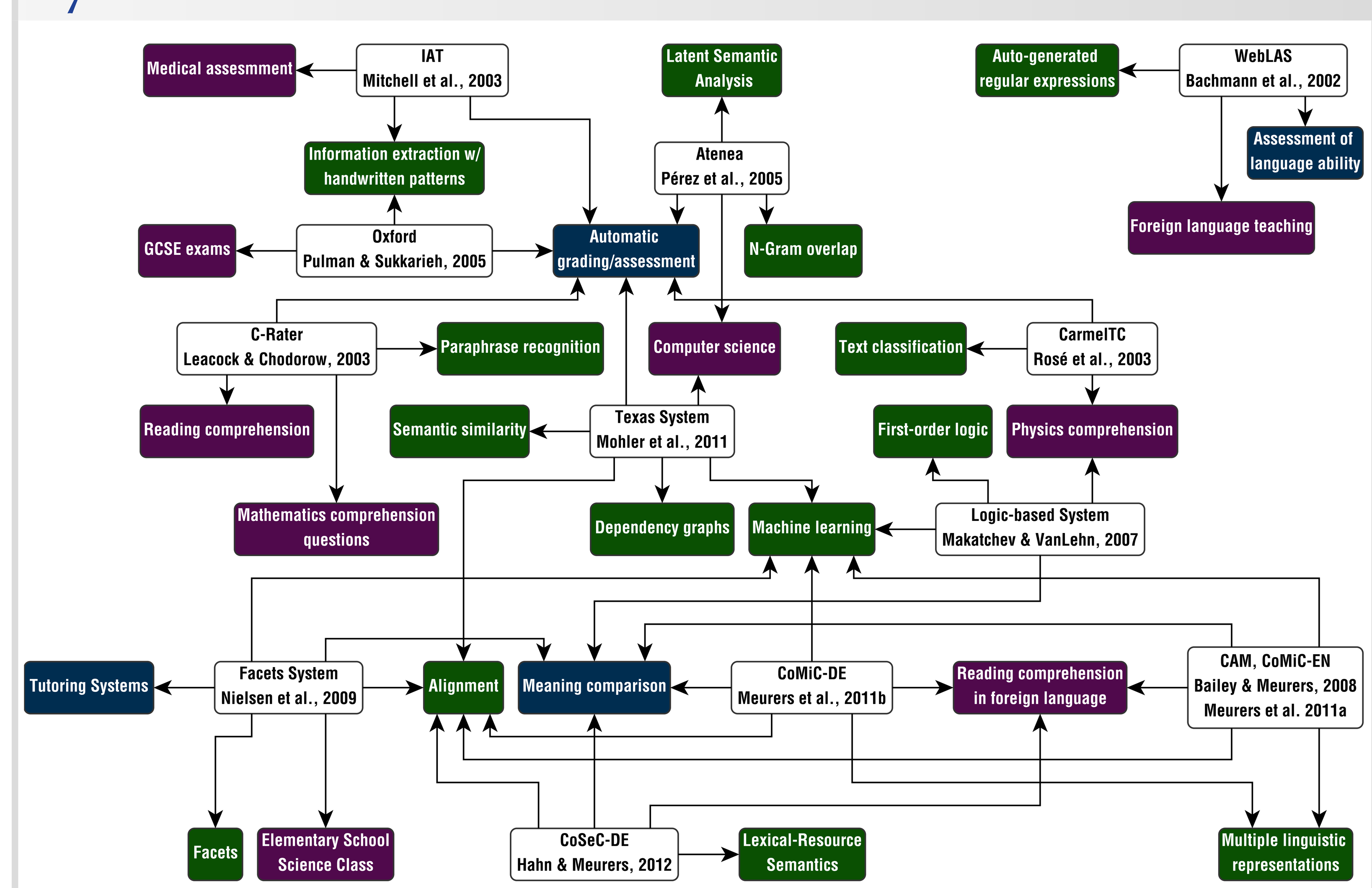
Gold Standard Ratings

- Low agreement for the two graders of Texas corpus (Mohler et al., 2011):
  - Pearson correlation (r) = 0.586
  - Root Mean Square Error (RMSE) = 0.659
- Should responses without perfect agreement be used in training and testing systems?
  - In other approaches, disagreements are resolved or the respective instances left out, cf., e.g., Beigman Klebanov & Beigman (2009).
  - In the Texas corpus, Mohler et al. (2011) opted to use the arithmetic mean of two raters as gold standard.
    But: Arithmetic mean is only reliable when using many raters.
- ➥ Meaningfulness of a gold standard for a task that humans cannot reliably perform needs attention. Can the task or the guidelines be improved?

The Short Answer Assessment Landscape



Systems



Legend: System | Goal | Domain | Technique

Comparing two Concrete Systems

Data (Mohler, Bunescu & Mihalcea, 2011)

- Corpus of 10 assignments and 2 exams from introductory CS class
- 2,442 student responses to 87 questions in total
  - avg. response length 18.4 tokens
- Each response rated by two human raters on 0–5 scale
  - exact grader agreement: 57.7%
  - gold standard created by averaging between raters
- Score distribution: Mean $\overline{x} = 4.19$, and Std. Deviation $s = 1.11$

Approaches

- Texas system (Mohler, Bunescu & Mihalcea, 2011)
  - Scoring system, using interval scale
  - Two components: Dependency Graph Alignment and Bag-of-word measures (e.g., LSA, tf∗idf)
  - SVR/SVMRank produces final numeric outcome based on features from the two components
- CoMiC-EN (Meurers, Ziai, Ott & Bailey, 2011a)
  - Meaning comparison system, using nominal scale
  - Annotation phase enriches input with linguistic information.
  - Alignment uses linguistic information to create mappings between student and target responses.
  - Classification (TiMBL) identifies meaning equivalence or nature of divergence from target based on 13 features from Alignment.

Evaluation

- CoMiC-EN not designed to perform scoring with numeric scales
  - ➥ Switch ML component from Memory-Based Learning to Support Vector Regression (SVR) using same feature set
- Setup as described by Mohler et al. (2011): 12-fold cross-validation SVR with linear kernel and tuned parameters based on training set
- Result: Texas system performs better on its own data

	Pearson Correlation	Root Mean Square Error
Mohler et al. (2011)	0.518	0.978
CoMiC-EN with SVR	0.405	1.016
Median Baseline		1.375

References

Bachman, L., N. Carr, G. Kamei, M. Kim, M. Pan, C. Salvador & Y. Sawaki (2002). A Reliable Approach to Automatic Assessment of Short Answer Free Responses. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*. pp. 1–4.

Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein & R. D. Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115.

Beigman Klebanov, B. & E. Beigman (2009). From annotator agreement to noise models. *Computational Linguistics* 35(4), 495–503.

Hahn, M. & D. Meurers (2012). Evaluating the Meaning of Answers to Reading Comprehension Questions: A Semantics-Based Approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*. Montreal.

Leacock, C. & M. Chodorow (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities* 37, 389–405.

Makatchev, M. & K. VanLehn (2007). Combining Baysian Networks and Formal Reasoning for Semantic Classification of Student Utterances. In *Proceedings of the International Conference on AI in Education (AIED)*. Los Angeles.

Meurers, D., R. Ziai, N. Ott & S. Bailey (2011a). 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.

Meurers, D., R. Ziai, N. Ott & J. Kopp (2011b). 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*. Edinburgh, Scotland, UK: ACL, pp. 1–9.

Mitchell, T., N. Aldrige & P. Broomhead (2003). Computerized Marking of Short-Answer Free-Text Responses. Paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Mohler, M., R. Bunescu & R. Mihalcea (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 752–762.

Nielsen, R. D., W. Ward & J. H. Martin (2009). Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering* 15(4), 479–501.

Pérez, D., E. Alfonseca, P. Rodríguez, A. Gliozzo, C. Strapparava & B. Magnini (2005). About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. *Revista signos* 38(59), 325–343.

Pulman, S. G. & J. Z. Sukkarieh (2005). Automatic Short Answer Marking. In J. Burstein & C. Leacock (eds.), *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 9–16.

Rosé, C. P., A. Roque, D. Bhembe & K. VanLehn (2003). A Hybrid Approach to Content Analysis for Automatic Essay Grading. In *Proceedings of HLT-NAACL 2003, short papers, Volume 2*. Edmonton, Canada: Association for Computational Linguistics, NAACL-Short '03, pp. 88–90.