Evaluating Answers to Reading Comprehension Questions in Context

Ramon Ziai[1], Niels Ott[1], Detmar Meurers[2]
[1]Sonderforschungsbereich 833, [2]Seminar für Sprachwissenschaft, Universität Tübingen

Research Questions

- What linguistic representations can be used robustly and efficiently in an automatic meaning comparison?
- What is the role of context and how can we utilize knowledge about it in comparing meaning automatically?
  Context here means questions and reading texts in reading comprehension tasks.

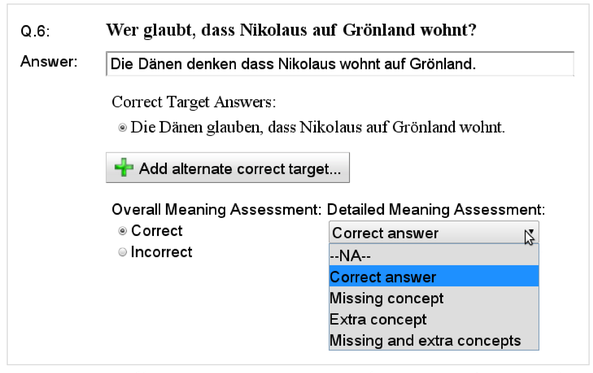
Why Reading Comprehension Exercises?

- Answers are different realizations of the same meaning.
- Meaning is clearly restricted by the task context (question, text).
- Learner language is not necessarily well-formed

➥ requires robust computational processing.

Corpus: CREG

- Corpus of Reading comprehension Exercises in German
- Is being collected in the German programs of the Ohio State University and Kansas University: almost only English L1.
- Meta data: information about students collected term by term.
- All learner answers are rated with respect to meaning (not form) by two annotators at the corresponding universities:



Assessment of learner answers in WELCOME (Meurers, Ott & Ziai, 2010)

- Agreement study based on a snapshot of the data from May 25, 2011 (Ott, Ziai & Meurers, to appear):

	# Student Answers	# Questions	% agreement binary	% agreement detailed	κ agreement binary	κ agreement detailed
KU:	5257	202	88.5%	86.6%	0.712	0.771
OSU:	4826	142	85.7%	70.6%	0.572	0.473

- Binary assessment is the observation of a task that teachers usually perform in grading: good percentage agreement.
- Detailed assessment: agreement drop in OSU data.
- Further research: have one team annotate a balanced subset of the data from the other team in order to level out effects of skewed category distribution ($\rightarrow$ low κ) in agreement study.

CoMiC: A Content Assessment System

Comparing Meaning in Context (CoMiC)

- CoMiC automatically judges whether or not a student answer is a correct answer to a reading comprehension question on basis of meaning comparison to a pre-defined target answer.
- CoMiC is a re-implementation and successor of the Content Assessment Module (CAM) by Bailey & Meurers (2008).

A Three-Phase Approach

1. Automatic Annotation enriches student and target answers as well as questions with information on different levels and types of abstraction.
2. Alignment maps elements of the learner answer to elements of the target response using annotation.
3. Classification analyzes the possible alignments and labels the learner response with a binary content assessment and a detailed diagnosis code.
   - Machine learning (TiMBL, Daelemans et al. 2007)

Example Alignment

Q: **Was sind die Kritikpunkte, die Leute über Hamburg äußern?**
'What are the objections people have about Hamburg?'

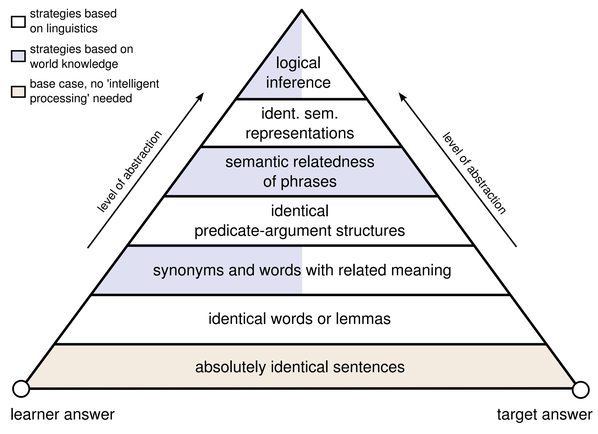
TA: **Der Gestank von Fisch und Schiffsdiesel an den Kais .**
The stink of fish and fuel on the quays .
SemType, Spelling, Spelling, Ident, Ident, Similarity

SA: **Der Geruch zon Fish und Schiffsdiesel beim Hafen .**
The smell of$_{err}$ fish$_{err}$ and fuel at the port .

Performance

- CoMiC-DE performs with an accuracy of 84,6% in an experiment with 1032 learner answers to 177 questions with 223 target answers (Meurers, Ziai, Ott & Kopp, 2011).
  - With correct and incorrect answers being equally distributed in the test data ($\rightarrow$ 50% random baseline).
- This is state-of-the-art compared to other systems for English, e. g., C-Rater (Leacock & Chodorow, 2003) or CAM (Bailey & Meurers, 2008)

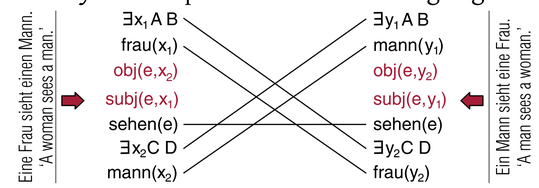
Levels of Abstraction in Meaning Comparison



Current Work in Progress

Semantic Representations in Meaning Comparison

- Basic idea: alignment (of parts) of PARTS-lists in Lexical Resource Semantics representations (LRS, Richter & Sailer 2004).
- Only 'structurally similar' parts of terms are being aligned:



(Hahn & Meurers, 2011)

Using Information Structure

- Focussed parts of answers encode requested information:

Q: **An was denken viele Menschen, wenn sie von Weißrussland hören?**
'What do many people think of when they hear about Belarus?'

TA: **Sie denken an die Tschernobyl-Katastrophe von 1986.**
'They think of the Chernobyl disaster of 1986.'

SA: **Ausländer denken bei Weißrussland weniger an Urlaub, sondern eher an die Tschernobyl-Katastrophe von 1986. Damals explodierten in der Sowjetunion Teile eines Atomkraftwerks und wurden einige Regionen Weißrusslands von der radioaktiven Strahlung verseucht.**
'Foreigners thinking about Belarus think less of vacation but rather of the Chernobyl disaster of 1986. Back then, parts of a nuclear plant exploded and some areas of Belarus were polluted by the radioactivity.'

- Distinguishing given and new information is not sufficient:

Q: **Ist die Wohnung in einem Neubau oder in einem Altbau?**
'Is the flat in a new building or in an old building?'

TA: **Die Wohnung ist in einem Neubau.**
'The flat is in a new building.'

SA: **Die Wohnung ist in einem Neubau**
'The flat is in a new building.'

Distributional Semantics of Phrasal Elements

- Semantic relatedness measures that use large corpora (e. g., PMI-IR, Turney 2001) are mostly used on the word level.
- Transfer of these approaches to phrases for non-compositional elements requiring world knowledge, e. g., *at home* vs. *in my house*.

References

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.

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch (2007). *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, P.O. Box 90153, NL-5000 LE, Tilburg, The Netherlands, version 6.0 ed.

Hahn, M. & D. Meurers (2011). On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona.

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

Meurers, D., N. Ott & R. Ziai (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Pre-Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217.

Meurers, D., R. Ziai, N. Ott & J. Kopp (2011). 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: Association for Computational Linguistics, pp. 1–9.

Ott, N., R. Ziai & D. Meurers (to appear). Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In T. Schmidt & K. Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam: Benjamins, Hamburg Studies in Multilingualism (HSM).

Richter, F. & M. Sailer (2004). Basic Concepts of Lexical Resource Semantics. In A. Beckmann & N. Preining (eds.), *European Summer School in Logic, Language and Information 2003. Course Material I*, Wien: Publication Series of the Kurt Gödel Society, vol. 5 of *Collegium Logicum*, pp. 87–143.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*. Freiburg, Germany, pp. 491–502.