Comparing Meaning in Context (CoMiC)

Detmar Meurers
Universität Tübingen

joint work with
Stacey Bailey, Niels Ott, Ramon Ziai

Université Paris 7, UFR d'Études Anglophone
March 25, 2011

Introduction

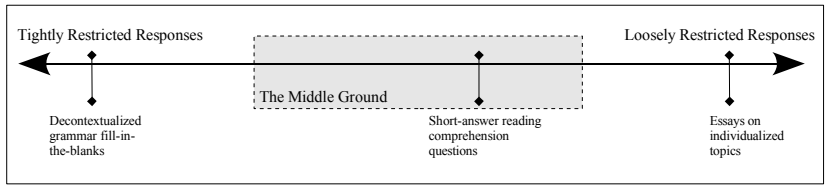
- Meaningful, contextualized use of language plays a crucial role in second language acquisition.
  - Yet automatic analysis has focused on form aspects, severely limiting applicability of ICALL tools in real-life.
- How can the meaning of sentences and text fragments be analyzed and compared in realistic situations?
  - Realistic situations:
    - differences in situative and world knowledge
    - language not necessarily well-formed
- Two challenges:
  - Which linguistic representations can be robustly identified as basis of a computational approximation of meaning?
  - How can the role of the task and context be integrated?

⇒ Start by collecting data of authentic language in context.

Collecting authentic data for content assessment

- Language is produced in concrete linguistic and extra-linguistic contexts.
  - This contextual setting includes situational knowledge and world knowledge.
- Current learner corpora typically consist of essay data, so only the essay topic is known.
  - contents quite unconstrained and not predictable
- We want to make the context explicit by collecting data in the setting of a concrete task.
  - For which tasks can content assessment be provided, without extensive world knowledge representation?

Exercise Spectrum & its Analysis Requirements



- Some activities in the middle ground provide explicit, language-based context with predictable contents:
  - summarization, reading comprehension, information-gap activities, . . .

⇒ Compile a corpus with answers to reading comprehension exercises written by learners of English (CREE) (Bailey 2008; Bailey & Meurers 2008; Meurers, Ziai, Ott & Bailey 2011).

- In the CoMiC project, we focus on learners of German (CREG) (Meurers, Ott & Ziai 2010).

Reading Comprehension (RC) Questions

- Most constrained: multiple choice
  - Example: *When was Mozart born?*
    a) 1756 b) 1796 c) 1812 d) 1917
- Least constrained: open-ended questions
  - Example: *How does the health system described in the text compare to that in your home country?*

⇒ Loosely restricted reading comprehension questions:
- It is possible to specify target answers.
- Responses can exhibit variation on lexical, morphological, syntactic, semantic levels.
- Common activity in real-life foreign language teaching.

Loosely restricted reading comprehension

An example

Question: *What are the methods of propaganda mentioned in the article?*

Target: *The methods include use of labels, visual images, and beautiful or famous people promoting the idea or product. Also used is linking the product to concepts that are admired or desired and to create the impression that everyone supports the product or idea.*

Sample Learner Responses:

- *A number of methods of propaganda are used in the media.*
- *Bositive or negative labels.*
- *Giving positive or negative labels. Using visual images. Having a beautiful or famous person to promote. Creating the impression that everyone supports the product or idea.*

Corpus of Reading comprehension Exercises in English (CREE)

- The corpus was collected in second language classrooms, using the ordinary exercises assigned by the teacher.
  - Teachers also provided target answers and learner answer assessment.
- CREE corpus: 566 responses to RC questions from intermediate English as a Second Language students.
  - Development set:
    - 311 responses from 11 students to 47 questions
  - Test set:
    - 255 responses from 15 students to 28 questions

Annotation: Categories for content assessment

- The annotation scheme was developed by analyzing target and learner responses in the development corpus.
- Two graders independently annotated the data:
  - detection (binary): correct vs. incorrect meaning
  - diagnosis (5 codes): correct; missing concept, extra concept, blend, non-answer

  Eliminated responses which graders did not agree on
  - 48 in development set (15%) and 31 in test set (12%)
- Learner responses vary significantly
  - no full bag-of-word overlap between test set answers and targets
- On average, 2.7 form errors per sentence.

Basic idea: Comparing responses & targets

Frage Where was Bob Hope when he heard about the news?

Musterantwort Bob Hope was at home.

Lernerantwort He was in his house.

From Annotation and Alignment to Diagnosis

- The Content-Assessment Module (CAM) compares target & learner responses in three steps:
  1. **Annotation** uses NLP to enrich the learner and target responses and question text with linguistic information.
  2. **Alignment** maps units in the learner response to units in the target response using the annotated information.
  3. **Diagnosis** analyzes the alignment to label the learner response with a target modification diagnosis code.

Types of alignment

Alignment can involve different types of representation:

Alignment Type	Example Match
Token-identical	**advertising** **advertising**
Lemma-resolved	**advertisement** **advertising**
Spelling-resolved	**campaing** **campaign**
Reference-resolved	**Clinton** **he**
Semantic similarity-resolved	**initial** **beginning**
Specialized expressions	**May 24, 2007** **5/24/2007**

Levels of alignment

Alignment can take place at different levels of representation:

Level	Example	Alignment
Tokens	**The explanation is simple.** **The reason is simple.**	**explanation** **reason**
Chunks	**A brown dog sat in a nice car.** **A nice dog sat in a car.**	**a brown dog** **a nice dog**
Dependency triples	**He knows the doctor.** **John knows him.**	**obj(knows, doctor)** **obj(knows, him)**

NLP tools used

Annotation Task	Language Processing Tool
Sentence Detection, Tokenization, Lemmatization	MontyLingua (Liu 2004)
Lemmatization	PC-KIMMO (Antworth 1993)
Spell Checking	Edit distance (Levenshtein 1966), SCOWL word list (Atkinson 2004)
Part-of-speech Tagging	TreeTagger (Schmid 1994)
Noun Phrase Chunking	CASS (Abney 1996)
Lexical Relations	WordNet (Miller 1995)
Similarity Scores	PMI-IR (Turney 2001; Mihalcea et al. 2006)
Dependency Relations	Stanford Parser (Klein & Manning 2003)

Features used for content assessment

- Diagnosis is based on 14 features:

  # of Overlapping Matches:
  - keyword (head word)
  - target/learner token
  - target/learner chunk
  - target/learner triple

  Nature of Matches:
  - % token matches
  - % lemma matches
  - % synonym matches
  - % similarity matches
  - % sem. type matches
  - match variety

  Semantic error detection

- For combining the evidence, machine learning (TiMBL, Daelemans et al. 2007) worked better than manual rules.

Results

Binary classification	Accuracy
Development Set (leave-one-out testing)	87%
Test Set	88%

Diagnosis with 5 codes	Accuracy
Development Set	87%
Test Set	87%

Form errors don't negatively impact results:
- 68% of **correctly** diagnosed items had form errors.
- 53% of **incorrectly** diagnosed ones did as well.

Related work

- No directly comparable systems, but competitive with results for automatic scoring of native speaker short answers by C-Rater (Leacock & Chodorow 2003; Leacock 2004).
  - Techniques used by essay grading systems (e.g., E-Rater, Burstein et al. 2003; AutoTutor, Graesser et al. 1999) do not generalize well to short (1-2 sentence) responses.
- Related research issues
  - Paraphrase recognition (e.g., Brockett & Dolan 2005; Hatzivassiloglou et al. 1999)
  - Machine translation evaluation (e.g., Banerjee & Lavie 2005; Lin & Och 2004)
  - Essay-based question answering systems (e.g., Deep Read, Hirschman et al. 1999)
  - Automatic grading (e.g., Leacock 2004; Marín 2004)
  - Recognition of Textual Entailment (RTE, Dagan et al. 2006)

Future work

Towards Interpretation in Context

- The reading comprehension question task we are focusing on provides an explicit context in form of
  - the text, and
  - the questions asked about it.
- CAM currently takes this context into account for basic anaphora resolution in the target and learner answers.
- But how about about other aspects of this context?
  - How should information in the answers that is *given* in the question be interpreted?
  - How can the nature of a question and the task strategies it requires be taken into account?

Towards interpretation in context

Treatment of *given* information

- Example from CREE:
  - **Cue**: *What was **the** major **moral question raised by the Clinton incident**?*
  - **Target**: ***The moral question raised by the Clinton incident was** whether a politician's person life is relevant to their job performance.*
  - **Response**: ***A basic question for the media is** whether a politician's personal life is relevant to his or her performance in the job.*

Information *given* in the question

Aspects of an approach

- The information in a response that is explicitly given in the question should not raise the number of matched units between target and learner answer.
- The original CAM simply removed *given* words.
- We are developing a more sophisticated approach to
  - keep sentence intact for deeper processing
  - use the occurrence of *given* information to distinguish between incorrect answers and off-topic answers.

Information *given* in the question

New information focus vs. contrastive focus

- Example based on CREG example pattern:
  - **Cue**: *Is Prague best known for its wine, beer, or coffee?*
  - **Target**: *The city is particularly well-known for its beer.*
  - **Response**: *Prague has a good reputation in terms of its excellent beer.*

→ Given information needs to be taken into account as answering a question for certain question types.

Towards interpretation in context

Question classification

- Comparing the meaning of answers to questions should make use of nature of the task and context.
- Features to be investigated include
  - **Learning Goals**: Targeted cognitive skills and knowledge (e.g., Anderson & Krathwohl 2001)
  - **Knowledge Sources**: The implicit/explicit answer source (Irwin 1986; Pearson & Johnson 1978)
  - **Text Type**: The rhetorical structure of the text (Champeau de Lopez et al. 1997)
  - **Answer Type**: The kind of answer expected (Gerbault 1999)
- Results here may also help answer:
  - What are suitable, more fine grained diagnosis categories for content assessment?

Diagnosis categories for comparing meaning

- Content assessment in the CAM currently distinguishes:
  - correct
  - missing concept
  - extra concept
  - blend
  - non-answer
- What are suitable and obtainable diagnosis categories for content assessment?
  - E.g., more detailed categories based on answer typing!

Adaptivity of analysis

Combining shallow and deep analysis

- Given the high number of form errors in learner data, deep analysis and model construction often is not feasible.
- However, there are patterns for which a dedicated, deep analysis may be possible or even important.
- Patterns to be explored include
  - semantic units expected in the answer (cf. answer typing)
  - specific linguistic constructions identified in the answer which require special treatment (e.g., negation).
  - typical well-formed “islands of compositionality” supporting a deep analysis (e.g., particular NP patterns)
- We intend to explore the identification of such patterns and how they can adaptively be integrated.

Beyond English

- Our original CAM work and related research topics (e.g., RTE) have generally focused on English.
- Adaptively combining shallow & deeper analyses becomes especially important when going from English to languages richer morphology & freer word order.
  - In the CoMiC project, we are compiling a large German reading comprehension corpus.

Conclusion

- A range of activities in current foreign language teaching practice support meaningful, contextualized interaction.
- Loosely restricted reading comprehension questions are an interesting activity type for exploring content assessment using adaptive, shallow-to-deeper content-analysis techniques.
- Machine learning can benefit shallow content assessment even for the small data sets typically available in ICALL research.
- Diagnosis results are comparable to detection results, but a larger corpus is needed for more detailed analysis.
- We identified a range of current research avenues: increased integration of context information, diagnosis categories for meaning comparison, effect of German morphology and word order on such an approach.

Publications related to CoMiC project

- CAM approach for English
  - (Bailey 2008; Bailey & Meurers 2008)
- Longitudinal corpus collection using WELCOME → KU/OSU collaboration
  - (Meurers, Ott & Ziai 2010)
- Dependency parsing of learner language
  - (Ott & Ziai 2010)
- Architecture for parallel analysis modules and CREE
  - (Meurers, Ziai, Ott & Bailey 2011)

References

Abney, S. (1996). Partial Parsing via Finite-State Cascades. In *The Robust Parsing Workshop of the European Summer School in Logic, Language and Information (ESSLLI '96)*. Prague, Czech Republic, pp. 1–8. URL http://www.vinartus.net/spa/97a.pdf.

Amaral, L., V. Metcalf & D. Meurers (2006). Language Awareness through Re-use of NLP Technology. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii. URL http://purl.org/net/icall/handouts/calico06-amaral-metcalf-meurers.pdf.

Anderson, L. W. & D. Krathwohl (eds.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman Publishers.

Antworth, E. L. (1993). Glossing Text with the PC-KIMMO Morphological Parser. *Computers and the Humanities* 26, 475–484. URL http://www.springerlink.com/content/r20w66k70976ur9l/fulltext.pdf.

Atkinson, K. (2004). Spell Checking Oriented Word Lists (SCOWL). URL http://wordlist.sourceforge.net/. Web resource.

Bailey, S. (2008). Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English as a Second Language. Ph.D. thesis, The Ohio State University. URL http://purl.org/net/Bailey-08.pdf.

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. URL http://aclweb.org/anthology/W08-0913.

Banerjee, S. & A. Lavie (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*. URL http://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf.

Brockett, C. & W. B. Dolan (2005). Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. pp. 1–8. URL http://aclweb.org/anthology/I05-5001.

Burstein, J. & M. Chodorow (1999). Automated Essay Scoring for Nonnative English Speakers. In *Proceedings of a Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, Joint Symposium of the Association of Computational Linguistics (ACL-99) and the International Association of Language Learning Technologies*. pp. 68–75. URL http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf.

Burstein, J., M. Chodorow & C. Leacock (2003). Criterion: Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-03)*. Acapulco, Mexico, pp. 3–10. URL http://ftp.ets.org/pub/res/erater_iaai03_burstein.pdf.

Champeau de Lopez, C., G. Marchi & M. Arreaza-Coyle (1997). A Taxonomy: Evaluating Reading Comprehension in EFL. *English Teaching Forum* 35(2), 30–42. URL http://dosfan.lib.uic.edu/usia/E-USIA/forum/vols/vol35/no2/p30.htm.

Daelemans, W., J. Zavrel, K. 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. URL http://ilk.uvt.nl/downloads/pub/papers/ilk.0703.pdf.

Dagan, I., O. Glickman & B. Magnini (2006). The PASCAL Recognising Textual Entailment Challenge. In J. Q. Candela, I. Dagan, B. Magnini & F. d'Alché Buc (eds.), *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*. Springer, vol. 3944 of *Lecture Notes in Computer Science*, pp. 177–190. URL http://u.cs.biu.ac.il/~dagan/publications/RTEChallenge.pdf.

Gerbault, J. (1999). Towards an analysis of answers to open-ended questions in computer-assisted language learning. In S. Lajoie & M. Vivet (eds.), *Proceedings of AIED*. IOS Press, pp. 686–689.

Graesser, A. C., K. Wiemer-Hastings, P. Wiemer-Hastings & R. Kreuz (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research* 1, 35–51.

Hatzivassiloglou, V., J. Klavans & E. Eskin (1999). Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Proceedings of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP'99)*. College Park, Maryland, pp. 203–212. URL http://aclweb.org/anthology/W99-0625.

Hirschman, L., M. Light, E. Breck & J. Burger (1999). Deep Read: A Reading Comprehension System. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*. College Park, Maryland, pp. 325–332. URL http://www.eecs.berkeley.edu/~nimar/readings/hirschman1999.pdf.

Irwin, J. W. (1986). *Teaching Reading Comprehension Processes*. Engelwood Cliffs, New Jersey: Prentice-Hall, Inc.

Klein, D. & C. D. Manning (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*. Sapporo, Japan, pp. 423–430. URL http://aclweb.org/anthology/P03-1054.

Leacock, C. (2004). Scoring Free-Responses Automatically: A Case Study of a Large-Scale Assessment. *Examens* 1(3). URL http://www.nocheating.org/Media/Research/pdf/erater_examens_leacock.pdf.

Leacock, C. & M. Chodorow (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities* 37, 389–405. URL http://www.ingentaconnect.com/content/klu/chum/2003/00000037/00000004/05144721?crawler=true.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10(8), 707–710. URL http://www.mendeley.com/research/binary-codes-capable-of-correcting-insertions-and-reversals/.

Lin, C.-Y. & F. J. Och (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. pp. 605–612. URL http://www.mt-archive.info/ACL-2004-Lin.pdf.

Liu, H. (2004). MontyLingua: An End-to-End Natural Language Processor with Common Sense. Software Website. Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts. URL http://web.media.mit.edu/~hugo/montylingua/.

Marín, D. R. P. (2004). Automatic Evaluation of Users' Short Essays by Using Statistical and Shallow Natural Language Processing Techniques. Master's thesis, Universidad Autónoma de Madrid. URL http://www.ii.uam.es/~dperez/tea.pdf.

Metcalf, V. & D. Meurers (2006). Generating Web-based English Preposition Exercises from Real-World Texts. URL http://purl.org/net/icall/handouts/eurocall06-metcalf-meurers.pdf. EUROCALL 2006. Granada, Spain. September 4–7, 2006.

Meurers, D., N. Ott & R. Ziai (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. URL http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Meurers, D., R. Ziai, N. Ott & S. Bailey (2011). Integrating Parallel Analysis Modules to Evaluate the Meaning of Answers to Reading Comprehension Questions. *IJCEELL Special Issue on Automatic Free-text Evaluation* URL http://purl.org/dm/papers/meurers-ziai-ott-bailey-11.html.

Mihalcea, R., C. Corley & C. Strapparava (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA: American Association for Artificial Intelligence (AAAI) Press, vol. 21(1), pp. 775–780. URL http://www.cse.unt.edu/~rada/papers/mihalcea.aaai06.pdf.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM* 38(11), 39–41. URL http://aclweb.org/anthology/H94-1111.

Ott, N. (2009). Information Retrieval for Language Learning: An Exploration of Text Difficulty Measures. ISCL master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany. URL http://drni.de/zap/ma-thesis.

Ott, N. & R. Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In M. Dickinson, K. Müürisep & M. Passarotti (eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*. vol. 9 of *NEALT Proceeding Series*, pp. 175–186. URL http://www.sfs.uni-tuebingen.de/~rziai/papers/Ott.Ziai-10.pdf.

Pearson, P. D. & D. Johnson (1978). *Teaching Reading Comprehension*. New York: Holt, Rinehart and Winston.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49. URL http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

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.

Wiemer-Hastings, P., K. Wiemer-Hastings & A. Graesser (1999). Improving an Intelligent Tutor's Comprehension of Students with Latent Semantic Analysis. In S. Lajoie & M. Vivet (eds.), *Artificial Intelligence in Education*, IOS Press, pp. 535–542. URL http://eprints.kfupm.edu.sa/45213/1/45213.pdf.