Diagnosing meaning errors in short answers to reading comprehension questions

Stacey Bailey
The Ohio State University

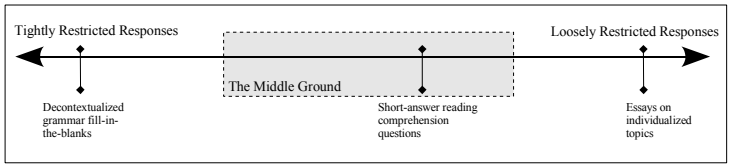
Detmar Meurers
Universität Tübingen

Workshop on Innovative Use of NLP
for Building Educational Applications
ACL 2008

June 19, 2008

Why care about meaning errors?

- Meaningful interaction in the foreign language is crucial for language learning.
- To be able to offer a wider range of activities, ICALL systems must be able to evaluate aspects of meaning.



- Loosely restricted reading comprehension (RC) questions are a good test case:
  - Common activity in real-life foreign language teaching.
  - Responses can exhibit variation on lexical, morphological, syntactic, semantic levels.
  - It is possible to specify target answers.

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.

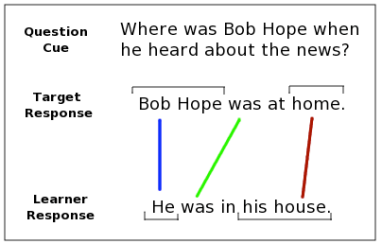
Our learner corpus

- Learner 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
- Teachers/graders provided target answers, keywords.
- Two graders 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%)
- On average, 2.7 form errors per sentence.
- Learner responses vary significantly; no full string or bag-of-word overlap with targets in test set.

Basic Idea: Comparing Responses and Targets

- Comparison at token, chunk and relation levels:



- Related research approaches similar tasks with many of the same techniques. This research includes
  - Automatic grading (e.g., Leacock 2004; Marín 2004)
  - Paraphrase recognition (e.g., Brockett and Dolan 2005; Hatzivassiloglou et al. 1999)
  - Machine translation evaluation (e.g., Banerjee and Lavie 2005; Lin and Och 2004)

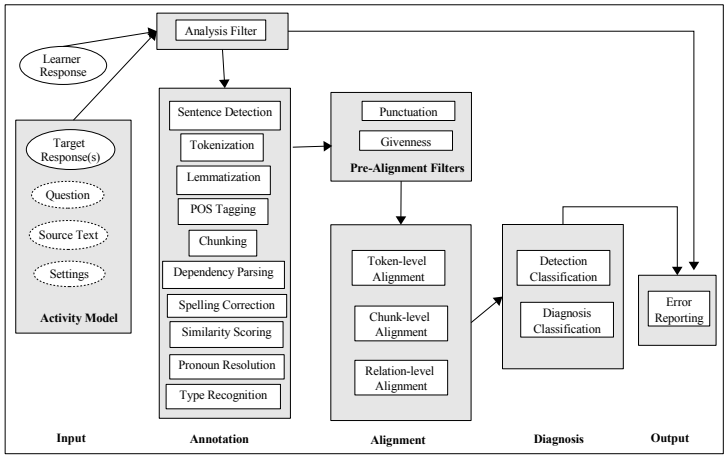
Content Assessment Module (CAM) Design

CAM compares target and learner responses in three phases:

1. Annotation uses NLP tools to enrich the learner and target responses, as well as the question text, with linguistic information, such as lemmas.
2. Alignment maps concepts in the learner response to concepts in the target response using the annotated information.
3. Diagnosis analyzes the alignment to label the learner response with a target modification diagnosis code.

The CAM Design

General Architecture



The CAM Design

NLP tools

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 and Manning 2003)

Error Diagnosis

- Diagnosis is based on 14 features:

  # of Overlapping Matches:
  - keyword (head)
  - 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

- We combined the evidence using
  - manual rules
    → did not generalize well from development to test set
  - machine learning (TiMBL, Daelemans et al. 2007), using majority voting on all distance measures

Results

Detection	Accuracy
Random Baseline	50%
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.

Results

Evaluation on a Balanced Set

- The development and test sets contain a high proportion of correct answers.
  - 71% of the development set and 84% of the test set were marked as correct by the human graders.
- We sampled a balanced set (50% correct answers), using all incorrect plus randomly selected correct ones.
  - balanced development set: 152 pairs
  - balanced test set: 72 pairs
- Accuracy results on balanced sets:
  - 50% random baseline
  - 78% on development set (leave-one-out testing)
  - 67% on test set

Related Work

- No directly comparable systems, but, e.g., results are competitive with accuracy reported for automatic scoring for native speaker short answers (Leacock 2004).
- ICALL systems typically
  - support exercise types that limit acceptable response variation and thus the need for sophisticated content assessment.
    e.g., German Tutor (Heift 2001), BANZAI (Nagata 2002)
  - restrict the topic domain and the nature of the input to be able to include deep content analysis.
    e.g., MILT (Kaplan et al. 1998), Herr Kommissar (DeSmedt 1995)
- Still other approaches focus on essays scoring
  e.g., E-rater (Burstein and Chodorow 1999), AutoTutor (Wiemer-Hastings et al. 1999)

Conclusion

- A range of activities in current foreign language teaching practice support meaningful, contextualized interaction.
- Taking loosely restricted reading comprehension questions as an example, we showed that content assessment for such activities is feasible using shallow 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.

References

Abney, Steven. 1996. Partial Parsing via Finite-State Cascades. *The Robust Parsing Workshop of the European Summer School in Logic, Language and Information (ESSLLI '96)*, 1–8. Prague, Czech Republic.

Antworth, Evan L. 1993. Glossing Text with the PC-KIMMO Morphological Parser. *Computers and the Humanities*, 26.475–484.

Atkinson, Kevin. 2004. Spell Checking Oriented Word Lists (SCOWL). Http://wordlist.sourceforge.net/.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, 65–72. Ann Arbor, Michigan.

Brockett, Chris and William B. Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 1–8. http://acl.ldc.upenn.edu/I/I05/I05-5001.pdf.

Burstein, Jill and Martin Chodorow. 1999. Automated Essay Scoring for Nonnative English Speakers. *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*, 68–75.

Diagnosing Meaning Errors in ICALL
Stacey Bailey
Detmar Meurers
Introduction
Why meaning errors?
Loosely restricted reading comprehension: An example
Our learner corpus
Basic idea behind approach
The CAM Design
General Architecture
NLP tools
Error Diagnosis
Results
Detection and Diagnosis Accuracy
Balancing the Test Set
Related Work
Conclusion
References
Appendix

Daelemans, Walter, Jakub Zavrel, Kovan der Sloot, and Antal 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 edition.

DeSmedt, William. 1995. Herr Kommissar: An ICALL Conversation Simulator for Intermediate German. V. Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology*, 153–174. Lawrence Erlbaum Associates.

Hatzivassiloglou, Vasileios, Judith Klavans, and Eleazar Eskin. 1999. Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. *Proceedings of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP'99)*, 203–212. College Park, Maryland.

Heift, Trude. 2001. Intelligent Language Tutoring Systems for Grammar Practice. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 6(2). http://www.spz.tu-darmstadt.de/projekt_ejournal/jg-06-2/beitrag/heift2.htm.

Kaplan, Jonathan, Mark Sobol, Robert Wisher, and Robert Seidel. 1998. The Military Language Tutor (MILT) Program: An Advanced Authoring System. *Computer Assisted Language Learning*, 11(3).265–287.

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

Leacock, Claudia. 2004. Scoring Free-Responses Automatically: A Case Study of a Large-Scale Assessment. *Examens*, 1(3).

Diagnosing Meaning Errors in ICALL
Stacey Bailey
Detmar Meurers
Introduction
Why meaning errors?
Loosely restricted reading comprehension: An example
Our learner corpus
Basic idea behind approach
The CAM Design
General Architecture
NLP tools
Error Diagnosis
Results
Detection and Diagnosis Accuracy
Balancing the Test Set
Related Work
Conclusion
References
Appendix

Levenshtein, Vladimir I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8).707–710.

Lin, Chin-Yew and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 605–612.

Liu, Hugo. 2004. MontyLingua: An End-to-End Natural Language Processor with Common Sense. http://web.media.mit.edu/~hugo/montylingua, accessed October 30, 2006.

Marín, Diana Rosario Pérez. 2004. *Automatic Evaluation of Users' Short Essays by Using Statistical and Shallow Natural Language Processing Techniques*. Master's thesis, Universidad Autónoma de Madrid. http://www.ii.uam.es/~dperez/tea.pdf.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *Proceedings of the National Conference on Artificial Intelligence*, volume 21(1), 775–780. Menlo Park, CA: American Association for Artificial Intelligence (AAAI) Press.

Miller, George. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11).39–41.

Nagata, Noriko. 2002. BANZAI: An Application of Natural Language Processing to Web-Based Language Learning. *CALICO Journal*, 19(3).583–599.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*, 44–49. Manchester, United Kingdom.

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

Diagnosing Meaning Errors in ICALL
Stacey Bailey
Detmar Meurers
Introduction
Why meaning errors?
Loosely restricted reading comprehension: An example
Our learner corpus
Basic idea behind approach
The CAM Design
General Architecture
NLP tools
Error Diagnosis
Results
Detection and Diagnosis Accuracy
Balancing the Test Set
Related Work
Conclusion
References
Appendix

Wiemer-Hastings, Peter, Katja Wiemer-Hastings, and Arthur Graesser. 1999. Improving an Intelligent Tutor's Comprehension of Students with Latent Semantic Analysis. Susanne Lajoie and Martial Vivet, editors, *Artificial Intelligence in Education*, 535–542. IOS Press.

Manual and Machine Learning Results

Detection	Accuracy
Baseline (random)	50%
Development Set: Manual CAM	81%
Development Set: CAM	87%
Test Set: Manual CAM	63%
Test Set: CAM	88%

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

Diagnosing Meaning Errors in ICALL
Stacey Bailey
Detmar Meurers