On automatically evaluating answers to reading comprehension questions

Stacey Bailey and Detmar Meurers

The Ohio State University

CALICO, San Marcos
May 24, 2007

Content Assessment in ICALL

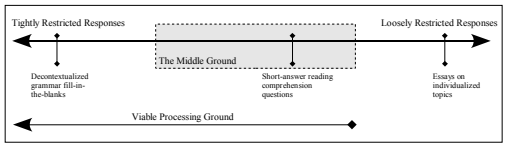
- Meaningful interaction in the foreign language is an important component of language learning.
- Can ICALL systems provide a range of meaning-based language activities?
  - To do so effectively, systems must be able to evaluate aspects of meaning of responses to those activities.
- We are calling this evaluation content assessment:
  - Analysis, diagnosis, and feedback regarding the appropriateness of the meaning in a learner response
- We are developing a Content Assessment Module (CAM)
  - to investigate the effectiveness of content assessment
  - using surface-level language processing strategies.

Situating CAM in ICALL

- Within the CALL domain, the majority of systems do not provide content assessment beyond string/token matching.
  - If the learner response is not identical to the target response, it is marked as incorrect.
- Existing ICALL systems (German-Tutor, Heift 2001; BANZAI, Nagata 2002) successfully avoid the need for sophisticated content assessment.
  - They control expected student input using specific activity types (e.g., build-a-sentence, translation).
- Other systems (e.g., Herr Kommissar, DeSmedt 1995) give the learner more linguistic freedom, but constrain the topic domain (e.g., interrogating suspects).

Language-learning Exercises

- Where can ICALL provide content assessment, without extensive linguistic and world knowledge representation?



- We focus on exercises in the middle ground.
  - If content assessment is possible there, it will be possible for more tightly restricted exercises too.
- As a test case, we target loosely restricted reading comprehension questions.

Reading Comprehension Questions

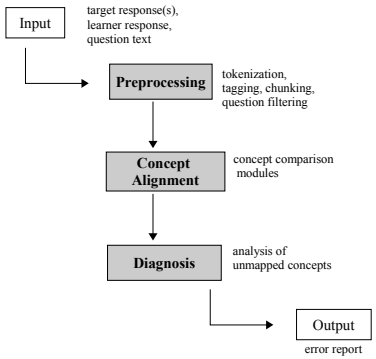
- Most constrained: multiple choice
  - Example: *When was Mozart born?*
    a) 1756 b) 1796 c) 1812 d) 1917
- Least constrained: open-ended questions
  - There is no right answer.
  - Evaluation is beyond current technology.
  - Example: *How do the statistics in your country compare to those in the text?*

⇒ Loosely restricted reading comprehension questions:
  - There is a correct answer.
  - The answer may be expressed in multiple ways.
  - Example: *What is alliteration?*

Designing a Content Assessment Module

- The CAM is intended as a component in a full ICALL system. (Stand-alone use is possible.)
- CAM compares a learner response to a target response and tries to “align” or match concepts.
  - Comparison relies on a series of modules to suggest different types of candidate alignments.
- Concepts that cannot be aligned by the modules should not be (i.e., no forced alignment).
- Diagnosis of errors requires accurate alignment.

CAM Architecture



Language Processing Tools in CAM

Task	Language Processing Tool
tokenization, lemmatization	MontyLingua (Liu 2004)
lemmatization	PCKimmo (Antworth 1993)
part-of-speech tagging	TreeTagger (Schmid 1994)
noun phrase chunking	CASS (Abney 1996)
lexical relations	WordNet (Miller 1995) WordNet::Similarity (Pedersen et al. 2004)
dependency relations	Minipar (Lin 1998)

Levels of Concept Alignment

Response concepts are aligned at different levels:

Level	Example	Alignment
Tokens	the explanation is simple compared to the reason is simple	explanation aligns to reason
Chunks	a good dog in a nice car compared to a nice dog in a fine car	a good dog aligns to a nice dog
Dependency triples	Rose knows the doctor compared to Rose knows John	obj(knows, doctor) aligns to obj(knows, John)

Surface-strategies for Concept Alignment

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

Challenges for Alignment
Issue 1: False Positives

- Token matching may incorrectly relate tokens.
  - **Target:** He opened a can of tuna.
    Response: You can go now.
- Removing morphological endings can remove relevant meaning distinctions.
  - **Target:** a fruitful discussion
    Response: a fruitless discussion
- Even alignments based on spelling errors can be nontrivial.
  - **Target:** He read the red book.
    Response: He raed the book.

Challenges for Alignment
Issue 2: Ill-formed Input

- Alignment may fail due to ill-formed learner responses.
  - Example: *Walk to a bar and says "Mak mine a double"*
- NLP tools are often designed for and trained on reasonably well-formed input (e.g., newspaper text).
  - A lack of data prevents retraining these tools.
- Candidate alignments may not always be reliable if based on incorrect analysis of ill-formed text.

Elements which don't need to be aligned
Information Given in the Question

- **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.*
- **Cue**: *What is alliteration?*
  - **Target**: *Alliteration is repetition of the initial sound in sequential words.*
  - **Response**: *where sequential words begin with the same letters or sounds*

Filtering Given Information

- The information in a response that is given in the question should not be evaluated.
- Responses should not be penalized or rewarded for the presence or absence of given text.
- Before learner and target responses are aligned, this information is filtered out using the question text.

Alignment Selection

- CAM collects and ranks all possible alignments between positions in the target and learner response.
- Currently, ranking is based on the type of mapping module used.
- Each type of alignment adds a weight to a final ranking score for each candidate.
- Based on the rankings, CAM selects the best alignment.

Diagnosis

- Based on unmapped concepts, CAM diagnoses possible meaning errors in responses.
- The targeted meaning errors are
  - lexical error
    - wrong word but related meaning
    - wrong word and unrelated meaning
  - argument placement error
  - missing concept
  - extra concept
  - other meaning error

Building a Learner Corpus

- Data collection began in Autumn 2006 to gather responses to reading comprehension questions.
- English learners from intermediate reading/writing courses of the American Language Program (ALP) at OSU provided data for the corpus.
- The ALP program offers intensive English instruction to undergraduate- and graduate-level students.
- Classes have less than 20 students, and learners are mostly Arabic, Chinese and Japanese speakers.

Data Overview

- The development corpus contains 311 responses to 47 questions from 11 students.
- Roughly another 300 responses are being collected and annotated as a test corpus.
- Instructors identified the target responses. They also provided correctness judgments for the test corpus.
- For the development set, there are 125 (40%) Incorrect responses, and 186 (60%) Correct responses.
  - It is often more challenging to process the Correct responses.

Examples of Collected Data

- **Cue:** *What is alliteration?*
  - **Target:** *Alliteration is repetition of the initial sound in sequential words.*
  - **Response:** *The worlds are often chosen to make some pattern or play on works. Sequential works begins with the same letter or sound.*
- **Cue:** *Which form of programming on TV shows the highest level of violence?*
  - **Target:** *Cartoons show the most violent acts.*
  - **Response:** *Television drama, children's programs and cartoons.*
- **Cue:** *Where was Bob Hope when he heard about the news?*
  - **Target:** *He was eating his breakfast at home.*
  - **Response:** *He was at home and eating his breakfast.*

Measuring Performance

Sample Responses	Count	CAM Judgments
Correct	120	67
Incorrect	120	173
Total	240	240

Performance Measure	Baseline	CAM
Precision	50%	67%
Recall	100%	97%

- CAM development is ongoing, but initial results for detection are promising.
- CAM detects an error in most cases it should but (still too) many that it shouldn't.

Ongoing Development

- **Alignment modules**: We are extending alignment to include more evidence for relating concepts.
- **Candidate ranking**: The current, hand-tuned scoring model will be replaced by a machine-learned one.
- **Diagnosis**: We will extend the module to examine number **and** types of unmapped elements.

More on Question Types

- Another extension to the study involves tying CAM performance to question features.
- The targeted reading comprehension questions are similar in terms of
  - the level of expected variation and
  - explicitness of their activity models.
- But such questions are not necessarily homogeneous.
- To tease apart question types that impact processing, we are investigating several features.

Question Classification Features

- Features investigated include
  - Learning Goals: Targeted cognitive skills and knowledge (e.g., Anderson and Krathwohl 2001)
  - Knowledge Sources: The implicit/explicit answer source (Irwin 1986; Pearson and 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)
- These features may prove useful for making better predictions about the effectiveness of the model.

Related Work beyond ICALL (1)

- Many of the alignment techniques in CAM have been used in non-ICALL systems.
  - Machine translation evaluation systems (e.g., METEOR, Banerjee and Lavie 2005)
  - Essay-based question answering systems (e.g., Deep Read, Hirschman et al. 1999
  - Summarization systems (e.g., Mani and Bloedorn 1997)
- These systems use surface-level matching for text comparisons, with varying degrees of success.
- Essay grading systems (e.g., E-Rater, Burstein et al. 2003) also use surface evaluation cues.
  - But these systems evaluate learner essays and not short (1-2 sentence) responses.

Related Work beyond ICALL (2)

- The closest system in function and purpose to CAM is C-Rater (Leacock and Chodorow 2003).
- As a grading system, it compares a student answer to a stored correct answer.
- After spelling correction and pronoun resolution, responses are compared using overlap/alignment of normalized dependency triples.
- C-Rater stops at detection (a score of 0, 1, or 2, where 1 is partial credit).
- On 100 student answers to 7 questions, C-rater obtained 84% agreement with human graders.
  - This is good for detection, but what about diagnosis?
  - Also, how would performance degrade on language-learner input?

Summary

- Our content assessment module
  - compares learner and target responses using shallow language processing strategies and
  - diagnoses errors by classifying unsuccessful concept matches.
- Development of CAM is guided by and tested using a corpus collected from English learners at OSU.
- This work makes the inclusion of reading comprehension activities feasible in an ICALL system.
- These activities
  - are widely used task-based activities and
  - require responses that exhibit linguistic variation.

  ⇒ A good test case for evaluating the model's effectiveness.

Outlook

- Ongoing work includes developing CAM to enhance alignment, ranking and diagnosis of learner errors.
- With the enhanced CAM, we will explore the relationship between performance and
  - alignment strategies,
  - language processing tools, and
  - question properties.

References

Abney, Steven. “Partial Parsing via Finite-State Cascades.” 1996.

Anderson, Lorin W., and David Krathwohl, editors. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives. New York: Longman Publishers, 2001.

Antworth, Evan L. “Glossing text with the PC-KIMMO morphological parser.” Computers and the Humanities 26: (1993) 475–484.

Banerjee, Satanjeev, and Alon Lavie. “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). 2005.

Burstein, Jill, Martin Chodorow, and Claudia Leacock. “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. 2003.

DeSmedt, William. “Herr Kommissar: An ICALL conversation simulator for intermediate German.” In Intelligent Language Tutors: Theory Shaping Technology, edited by V. Melissa Holland, Jonathan Kaplan, and Michelle Sams, Lawrence Erlbaum Associates, 1995, 153–174.

Gerbault, Jeannine. “Towards an analysis of answers to open-ended questions in computer-assisted language learning.” In Artificial Intelligence in Education, edited by Susanne Lajoie, and Martial Vivet, IOS Press, 1999, 686–689.

Content Assessment in ICALL
Stacey Bailey and Detmar Meurers
Introduction
Exercises Types
CAM
Design Idea
Architecture
Language Processing Tools
CAM Strategies
Alignment Levels
Alignment Types
Alignment Issues
Given Information
Alignment Selection
Response Diagnosis
Testing the Approach
Building a Corpus
Data Examples
Measuring Performance
Ongoing Development
Related Work
Summary and Outlook
References

Heift, Trude. “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.

Hirschman, Lynette, Marc Light, Eric Breck, and John Burger. “Deep Read: A Reading Comprehension System.” In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. 1999. URL http://citeseer.ist.psu.edu/hirschman99deep.html.

Irwin, Judith Westphal. Teaching Reading Comprehension Processes. Engelwood Cliffs, New Jersey: Prentice-Hall, Inc., 1986.

Leacock, Claudia, and Martin Chodorow. “C-rater: Automated Scoring of Short-Answer Questions.” Computers and the Humanities 37: (2003) 389–405.

Lin, Dekang. “Dependency-based Evaluation of MINIPAR.” In Workshop on the Evaluation of Parsing Systems. 1998.

Liu, Hugo. “MontyLingua: An end-to-end natural language processor with common sense.”, 2004. URL http://web.media.mit.edu/~hugo/montylingua. Last Accessed October 30, 2006.

Champeau de Lopez, Cheryl, Giancarla Marchi, and Maria Arreaza-Coyle. “A Taxonomy: Evaluating Reading Comprehension in EFL.” English Teaching Forum 35(2): (1997) 30–42.

Mani, Inderjeet, and Eric Bloedorn. “Multi-document summarization by graph search and matching.” In Proceedings of AAAI-97. 1997, 622–628.

Miller, George. “WordNet: a lexical database for English.” Communications of the ACM 38(11): (1995) 39–41.

Nagata, Noriko. “BANZAI: An application of natural language processing to Web based language learning.” CALICO Journal 19(3): (2002) 583–599.

27 / 27

Pearson, P. David, and Dale Johnson. Teaching Reading Comprehension. New York: Holt, Rinehart and Winston, 1978.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. “WordNet::Similarity - Measuring the Relatedness of Concepts.” 2004.

Schmid, Helmut. “Probabilistic Part-of-Speech Tagging Using Decision Trees.” 1994.

Content Assessment in ICALL
Stacey Bailey and Detmar Meurers
Introduction
Exercises Types
CAM
Design Idea
Architecture
Language Processing Tools
CAM Strategies
Alignment Levels
Alignment Types
Alignment Issues
Given Information
Alignment Selection
Response Diagnosis
Testing the Approach
Building a Corpus
Data Examples
Measuring Performance
Ongoing Development
Related Work
Summary and Outlook
References

27 / 27