Diagnosing meaning errors in Intelligent Computer-Assisted Language Learning

Walt Detmar Meurers
Universität Tübingen

joint work with Stacey Bailey

Nancy
January 29, 2009

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Introduction

- Computers widely used in foreign language teaching to help learners experience a foreign language & culture.
  - multimedia presentations, web-based TV/radio/news, email/chat with native speakers, . . .
- Apart from the undisputed role of contextualized, communicative language use, which other aspects are important for language acquisition?
- Research since the 90s has shown that **awareness of language forms and rules** is important for an adult learner to successfully acquire a foreign language.
  - (cf., e.g., Long 1991, 1996; Ellis 1994; Schmidt 1995; Lyster 1998; Lightbown and Spada 1999; Norris and Ortega 2000)

Linguistics and NLP for ICALL

- Linguistic analysis and NLP technology can be used in Computer-Aided Language Learning tools that
  - foster learner awareness of language forms & categories.
  - provide individual feedback on learner errors.
- Very few ICALL systems are used in FLT practice today (Nagata 2002; Heift 2001; Amaral and Meurers 2006).
- Problem: lack of interdisciplinary research combining computational, linguistic, and FLT/SLA expertise.
- Our general approach:
  - Link CL research to genuine FLT needs.

Current research foci in the OSU ICALL group

- TAGARELA System for Portuguese, an intelligent web-based workbook integrated in the OSU Portuguese Language Program (Amaral and Meurers 2005, 2006)
- Working with English Real Texts (WERTi)
  - Language awareness activities using enhanced real-life texts (Amaral et al. 2006; Metcalf and Meurers 2006)
  - Improved detection of word order errors (Metcalf and Meurers 2006)
- Content Assessment Module (Bailey and Meurers 2006)
  - Evaluation of semantic content of learner response by comparing the learner response to target responses.

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 working on a Content Assessment Module (CAM)
  - exploring contexts in which content assessment can be effective,
  - adaptively combining language processing strategies, from shallow to deeper analysis.

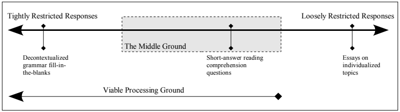
Situating CAM in ICALL

- 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; MILT, Kaplan et al. 1998) restrict the topic domain (e.g., interrogating suspects), and thereby the form of the input, to be able to include deep content analysis.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Language-learning Exercises

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



- We focus on exercises in the middle ground: loosely restricted reading comprehension questions.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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
  - 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:
  - 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.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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
- The corpus was collected in an ordinary second language classroom, using the questions and answers independently assigned by the teacher.
- Teachers/graders provided target answers and sometimes also target keywords.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Annotation: Categories for content assessment

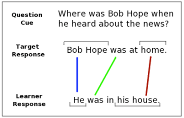
- The annotation scheme was developed by analyzing target and learner responses in the development corpus.
- This annotation scheme
  - focuses on how the learner response varies from target,
  - but assumes the learner is trying to "hit" the target(s).
- Two graders independently annotated the data:
  - detection (binary): correct vs. incorrect meaning
  - diagnosis (5 codes): correct; missing concept, extra concept, blend, non-answer
    - Also subclassified correct learner answers into those in line with target and those which are alternate answers.

  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.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Basic Idea: Comparing Responses and Targets

- Comparison at token, chunk and relation levels:



- Related research issues:
  - 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)
  - 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)

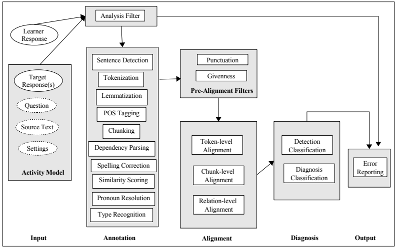
Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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 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.

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)

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	Rose knows the doctor. Rose knows him.	obj(knows, doctor) obj(knows, him)

Error Diagnosis Features

- Diagnosis is based on 14 features:

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

  Semantic error detection

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

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Combining the Evidence

- Explored combining the evidence using manual rules:

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

$\Rightarrow$ The manual rules do not generalize well from development to test set.

- We then used machine learning (TiMBL, Daelemans et al. 2007), with majority voting on all distance measures.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Related Work

- No directly comparable systems, but results are competitive with accuracy reported for automatic scoring for native speaker short answers (C-Rater, Leacock and Chodorow 2003; Leacock 2004).
  - C-rater performs diagnosis with three categories
  - Performance degradation on language-learner input?
- Essay grading systems (e.g., E-Rater, Burstein and Chodorow 1999; Burstein et al. 2003, AutoTutor Wiemer-Hastings et al. 1999).
  - Such systems evaluate learner essays and the techniques used do not generalize well to short (1-2 sentence) responses.

Towards Interpretation in Context

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

- The Recognizing Textual Entailment task has been pointed out be problematic in lacking a context in which the evaluation takes place (e.g., Manning 2006).
- The reading comprehension question task we are focusing on provides an explicit context in form of
  - the text, and
  - the question asked about it (i.e. the task).
- CAM currently takes this context into account for basic anaphora resolution for elements 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?
  - What is the nature of the questions and which task strategies do they require?

Information given in the question

Examples

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

- **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

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

- 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 current CAM version simply removes words included in both the question and the target and learner answers.
- A more sophisticated approach is needed to
  - keep all elements needed for deeper processing (e.g., parsing into dependency triples)
  - use the occurrence of *given* information to distinguish between partially incorrect answers (additional/missing units) and non-answers (totally missing the topic).

Question Classification

Motivation

- Another extension we are exploring takes a closer look at the nature of the questions.
- The targeted reading comprehension questions are similar in terms of
  - the level of expected variation and
  - explicitness of their activity models (target answer).
- But such questions are not necessarily homogeneous.
- To tease apart question types that impact processing, we are investigating several features.

Question Classification

Potentially relevant features

- Features to be 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)

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Adaptivity of analysis

Combining shallow and deep analysis

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

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Beyond English

- Our work and related research topics (e.g., RTE) have generally focused on English.
- How do content-assessment methods need to be adapted for a language with richer morphology and freer word order, such as German?

Conclusion

- NLP can be used in Computer-Aided Language Learning to provide individualized feedback and foster learner awareness of language forms & categories.
- To support meaningful, contextualized language learning tasks, automatic content assessment is crucial.
- Loosely restricted reading comprehension questions are an interesting activity type for exploring content assessment.
- Machine learning can be beneficial 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.
- Avenues for future research: use task and context information, better diagnosis categories for meaning comparison, adaptive combination of shallow and deep processsing, consider languages other than English.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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.

Amaral, Luiz, Vanessa Metcalf, and Detmar 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.

Amaral, Luiz and Detmar Meurers. 2005. Towards Bridging the Gap between the Needs of Foreign Language Teaching and NLP in ICALL. AntÛnio Pedros-Gascon, editor, *Proceedings of the 8th Annual Symposium on Hispanic and Luso-Brazilian Literatures, Linguistics, and Cultures*.

Amaral, Luiz and Detmar Meurers. 2006. Where does ICALL Fit into Foreign Language Teaching? 23rd Annual Conference of the Computer Assisted Language Instruction Consortium (CALICO), May 19, 2006. University of Hawaii.

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

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/.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Bailey, Stacey and Detmar Meurers. 2006. Exercise-driven selection of content matching methodologies. EUROCALL Conference. September 6, 2006. University of Granada.

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.

Burstein, Jill, Martin Chodorow, and Claudia Leacock. 2003. Criterion: Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-03)*, 3–10. Acapulco, Mexico.

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

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *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*, volume 3944 of *Lecture Notes in Computer Science*, 177–190. Springer.

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.

Ellis, Nick. 1994. Implicit and Explicit Language Learning - An Overview. *Implicit and Explicit Learning of Languages*, 1–31. San Diego, CA: Academic Press.

Gerbault, Jeannine. 1999. Towards an Analysis of Answers to Open-ended Questions in Computer-assisted Language Learning. Susanne Lajoie and Martial Vivet, editors, *Artificial Intelligence in Education*, 686–689. IOS Press.

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.

Hirschman, Lynette, Marc Light, Eric Breck, and John Burger. 1999. Deep Read: A Reading Comprehension System. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 325–332. College Park, Maryland. http://citeseer.ist.psu.edu/hirschman99deep.html.

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

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).

Leacock, Claudia and Martin Chodorow. 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, 37.389–405.

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

Lightbown, P. M. and N Spada. 1999. *How languages are learned.* Oxford: Oxford University Press.

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen
30 / 30

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.

Long, M. H. 1991. Focus on form: A design feature in language teaching methodology. K. De Bot, C. Kramsch, and R. Ginsberg, editors, *Foreign language research in cross-cultural perspective*, 39–52. Amsterdam: John Benjamins.

Long, Michael. 1996. The Role of the Linguistic Environment in Second Language Acquisition. W.C. Ritchie and T.K. Bhatia, editors, *Handbook of Second Language Acquisition*, 413–468. Academic Press.

Lyster, R. 1998. Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classroom. *Language Learning*, 48.183–218.

Manning, Christopher D. 2006. Local Textual Inference: It's Hard to Circumscribe, But You Know it When You See it – And NLP Needs It. Manuscript. Stanford University.

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.

Metcalf, Vanessa and Detmar Meurers. 2006. When to Use Deep Processing and When Not To – The Example of Word Order Errors. Pre-conference Workshop

Diagnosing Meaning Errors in ICALL
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen
30 / 30

on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii.

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.

Norris, J. and L. Ortega. 2000. Effectivenss of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis. *Language Learning*, 50(3).

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

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

Schmidt, R. 1995. Consciousness and foreign language: A tutorial on the role of attention and awareness in learning. R. Schmidt, editor, *Attention and awareness in foreign language learning*, 1–63. Honolulu: University of Hawaii Press.

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
Detmar Meurers
Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen
30 / 30

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.

Introduction
Meaning in ICALL
Exercises Types
Loosely restricted reading comprehension: An example
Our learner corpus
Annotation: Categories for content assessment
Basic idea behind approach
CAM Design
General Architecture
NLP tools
Alignment Types
Alignment Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Given inform. in question
Question Classification
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion