Intelligent Computer-Assisted Language Learning

Part III: Diagnosing Meaning Errors in ICALL

Detmar Meurers
(Universität Tübingen)

based on joint research with Stacey Bailey. See: Bailey/Meurers (2009): "Diagnosing Meaning Errors in Short Answers to Reading Comprehension". *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications.* http://purl.org/dm/papers/bailey-meurers-08.html

European Summer School in Language, Logic, and Information
Bordeaux. July 27–31, 2009

The importance of meaning

- Meaningful interaction in the foreign language is an essential component of second language acquisition.
  - Communicative language teaching, content-based instruction and task-based language teaching all stress the importance of meaning and exchange of information in language learning (Richards & Rodgers 2001).

⇒ Meaning (content) assessment is a critical component for intelligent computer-aided language learning (ICALL) systems in real-life language teaching practice.

ICALL: Part III Diagnosing meaning errors
Detmar Meurers Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Implications for ICALL activities

For an ICALL system to be effectively integrated into language instruction, it must

- offer more than drills and other form-based activities,
- provide a range of contextualized, meaningful language learning activities, and
- recognize multiple realizations of the same semantic content in learner responses to an activity.

ICALL: Part III Diagnosing meaning errors
Detmar Meurers Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Implications for ICALL content processing

An ICALL system that can be effectively integrated into different types of language instruction is one that is

- **Holistic**: The ICALL system should process both form and meaning of learner responses and, in the latter case, extract a representation of meaning,
- **Flexible**: Processing of learner responses must be adaptable, based on the goals of the activity.
- **Robust**: The system must analyze meaning even in the presence of form errors.

ICALL: Part III Diagnosing meaning errors
Detmar Meurers Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Existing ICALL systems

Background

- Until recently, research into morphological and structural processing has dominated NLP technology development.
- In consequence, most existing ICALL systems have addressed form assessment rather than meaning assessment.
- This emphasis on form assessment has limited the types of exercises that have been offered in existing ICALL systems.
  - German Tutor (Heift 2001) – Uses activities such as build-a-sentence that restricts responses to include supplied word forms.
  - BANZAI (Nagata 2002) – Extensively uses translation to restrict expected responses.

Eberhard Karls Universität Tübingen

Existing ICALL systems

Limitations

- Meaning assessment in existing ICALL systems is typically accomplished through form comparison.
  - If the form matches in comparing a learner and target response, the meaning is correct.
- This approach is successful due to restrictions on exercise types in which variation is not expected or allowed (Ex: cloze, build-a-sentence, translation).
- This limited processing fails for meaning assessment whenever variation occurs. For example:
  - Character-by-character string matching fails on responses with variation in capitalization or spacing.
  - Token-by-token string matching fails on variation in spelling, lexical material, word order or structure.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Eberhard Karls Universität Tübingen

Shifting the perspective of ICALL system design

- Fortunately, the analysis of meaning is increasingly a topic addressed by research in computational linguistics.
- It is possible to focus on what language instructors need – form or meaning processing – and to allow language exercises to drive the technology used in ICALL systems.
- To do this, we need to know
  - what existing language learning exercises should be targeted and what their properties are,
  - whether these exercises can be adapted to an ICALL system, and
  - whether existing NLP technology can effectively process the targeted exercise types.

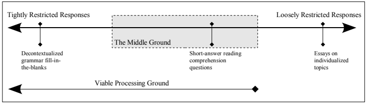
ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Eberhard Karls Universität Tübingen

Relating language exercises and NLP

- The more variation is possible in learner responses to an exercise, the more processing is required for meaning assessment.
- A spectrum of exercises and meaning analyses falls out of this relationship between exercises and NLP.



  - At one extreme, there are restricted exercise types requiring minimal analysis in order to assess meaning.
  - At the other extreme are free-response exercises requiring extensive meaning analysis and world knowledge.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Eberhard Karls Universität Tübingen

Exercise properties and content processing

1. **Level of expected response variation** – Lexical, morphological, structural, etc.
2. **Response length** – Multiple choice, single-word, phrase, sentence, paragraph, essay.
3. **Activity structure** – How much instruction is given about the intended form/meaning of the response.
4. **Target response** – Whether there is a specific correct answer that is clearly defined in the activity model.
5. **Assessment criteria** – What the goals of assessment are for the particular activity.

Exercise example 1

Guided fill-in-the-blank

Activity from Azar (1999), a grammar textbook for learners of Am. English:

> **Directions:** Complete the sentences with **no** or **not**:
>
> 1. *I can do it by myself. I need ____ help.*

- Many cloze exercises are designed for evaluating grammar skills (Ex: conjugation) and lexical choice.
- Little or no response variation is expected.
- There are only a finite number of target responses.
- To process meaning, a target may be stored and its form matched against that of the learner response.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Exercise example 2

Open-ended questions

Activity from Kirn & Hartmann (2002), a textbook for learners of English:

> **Directions:** In small groups, talk about your answers to these questions about your country.
>
> 1. How has technology changed the way in which people live and work?

- There is no specific expected target response; there is a wide range of possible answers of different lengths.
- Structural, morphological and lexical choice within that range may be highly variable.
- To extract and compare meaning, extensive linguistic knowledge, real-world knowledge, and NLP beyond the current technology is required.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

The middle ground

- The space between the opposite ends of the spectrum seems to offer good opportunities for combining real FLT needs with realistic computational processing and resources.
- The degree to which exercises in the middle ground can be easily, effectively and reliably processed with NLP technology is what we are exploring.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

A subset of exercises in the middle ground

- The focus of our research is on exercises with
  - clearly defined target responses and
  - expected variation in lexical, morphological and syntactic forms.
- The activities
  - represent common types of task-based activities in current approaches to language instruction,
  - emphasize meaning (comprehension and production),
  - support a range of assessment types, and
  - adapt easily to an ICALL setting.

Exemplifying the middle ground
Summarization

Activity from Seal (1997), a textbook for learners of English:

> **Directions:** Write a summary of the article "Coping with Stress." Remember to include only the main ideas and to omit highly specific details or supporting evidence.

- Summarization activities focus on the comprehension and reproduction of the essential meaning components of a text.
- Learner responses may be highly variable, but predictable given that the source text is known.
- Given a model summary, the learner response can be compared to the target model to evaluate its content.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Exemplifying the middle ground
Question answering

Activity from Seal (1997):

> **Directions:** Answer the following questions about the reading "Early Adulthood":
>
> 1. Why does the writer state that the factors that may influence an individual in the choice of a career may be "conflicting"?

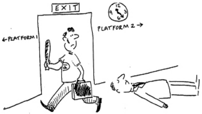
- Question answering activities often evaluate reading comprehension.
- Thus, target responses come directly from the source text.
- Again, learner responses may be highly variable, but a clearly definable target response to each question makes meaning assessment possible.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Exemplifying the middle ground
Information gap

Activity from Birch (2005):

> You will be asked questions…
>
> *About the robber:*
>
> Male or female, age, clothes, appearance, weapon
>
> *About the robbery:*
>
> Time, things stolen



- The activity design limits the range of acceptable target responses.
- Thus, the target content is suitably restricted, while the form of learner responses may be highly variable.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
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.

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.*

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Annotation: Categories for content assessment

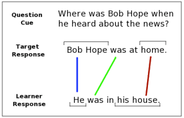
- The annotation scheme was developed by analyzing target and learner responses in the development corpus.
- This annotation scheme
  - focuses on how the learner response varies from target,
  - but assumes the learner is trying to “hit” the target(s).
- Two graders independently annotated the data:
  - 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.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
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 & 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)

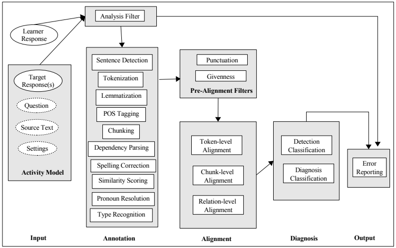
Content Assessment Module (CAM) Design

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 & 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)

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Error Diagnosis Features

- 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

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
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%

⇒ 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.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen
Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
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.

Related Work

- No directly comparable systems, but results are competitive with accuracy reported for automatic scoring for native speaker short answers (C-Rater, Leacock & 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 & Chodorow 1999; Burstein et al. 2003, AutoTutor Graesser et al. 1999; Wiemer-Hastings et al. 1999).
  - Such systems evaluate learner essays and the techniques used do not generalize well to short (1-2 sentence) responses.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Towards Interpretation in Context

- 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?

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

Information given in the question

Examples

- **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.*

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion
Eberhard Karls Universität Tübingen

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 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 & 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)

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, 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.

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?

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Eberhard Karls Universität Tübingen

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.
- CAM prototype (Bailey & Meurers 2008) shows that content assessment for such activities is feasible
- Avenues for future research: use task and context information, better diagnosis categories for meaning comparison, adaptive combination of shallow and deep processing, consider languages other than English.
  - ⇒ New SFB 833 Project A4 (2009–2013, with Niels Ott, Ramon Ziai): *Comparing Meaning in Context: Components of a shallow semantic analysis*

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Eberhard Karls Universität Tübingen

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.

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.

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

Azar, B. S. (1999). *Understanding and using English grammar*. Pearson Education, 3rd ed.

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, held at ACL 2008*. Columbus, Ohio: Association for Computational Linguistics, pp. 107–115. URL http://aclweb.org/anthology-new/W/W08/W08-0913.pdf.

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.

ICALL: Part III
Diagnosing meaning errors
Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective
Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach
CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English
Conclusion

Eberhard Karls Universität Tübingen

Birch, G. (2005). Balancing fluency, accuracy and complexity. In C. Edwards & J. Willis (eds.), *Teachers Exploring Tasks in English Language Teaching*, Palgrave Macmillan, pp. 228–239.

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.pdf.

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.

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.

Champeau de Lopez, C., G. Marchi & M. Arreaza-Coyle (1997). A Taxonomy: Evaluating Reading Comprehension in EFL. *English Teaching Forum* 35(2), 30–42.

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.

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.

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.

Heift, T. (2001). Error-Specific and Individualized Feedback in a Web-based Language Tutoring System: Do They Read It? *ReCALL* 13(2), 129–142.

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. http://citeseer.ist.psu.edu/hirschman99deep.html.

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

Kirn, E. & P. Hartmann (2002). *Interactions 2: Reading*. McGraw-Hill Contemporary, fourth ed.

ICALL: Part III Diagnosing meaning errors

Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective

Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach

CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English

Conclusion

Eberhard Karls Universität Tübingen

39 / 39

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).

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

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

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.

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

Manning, C. D. (2006). Local Textual Inference: It's hard to circumscribe, but you know it when you see it – and NLP needs it. URL http://nlp.stanford.edu/7Emanning/papers/LocalTextualInference.pdf. Ms. Stanford University.

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. http://www.ii.uam.es/˜dperez/tea.pdf.

Mihalcea, R., C. Corley & C. Strapparava (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the*

ICALL: Part III Diagnosing meaning errors

Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective

Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach

CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English

Conclusion

Eberhard Karls Universität Tübingen

39 / 39

National Conference on Artificial Intelligence. Menlo Park, CA: American Association for Artificial Intelligence (AAAI) Press, vol. 21(1), pp. 775–780.

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

Nagata, N. (2002). BANZAI: An Application of Natural Language Processingto Web based Language Learning. *CALICO Journal* 19(3), 583–599.

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

Richards, J. & T. Rodgers (2001). *Approaches and Methods in Language Teaching*. Cambridge University Press, second ed.

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. URL http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Seal, B. (1997). *Academic Encounters, Reading, Study Skills and Writing: Human Behavior*. Cambridge University Press.

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.

ICALL: Part III Diagnosing meaning errors

Detmar Meurers
Universität Tübingen

Introduction
Importance of Meaning
Existing ICALL / Limitations
Shifting the perspective

Exercise Spectrum
Exercise Properties
Examples
The Middle Ground
Reading comprehension
Our learner corpus
Gold standard annotation
Basic idea behind approach

CAM
General Architecture
NLP tools
Alignment Types & Levels
Error Diagnosis Features
Results
Related Work
Future work
Interpretation in Context
Diagnosis categories
Adaptivity (shallow/deep)
Beyond English

Conclusion

Eberhard Karls Universität Tübingen

39 / 39