Diagnosing meaning errors in short answers to reading comprehension questions

Stacey Bailey
Department of Linguistics
The Ohio State University
1712 Neil Avenue
Columbus, Ohio 43210, USA
s.bailey@ling.osu.edu

Detmar Meurers
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstrasse 19
72074 Tübingen, Germany
dm@sfs.uni-tuebingen.de

Abstract

A common focus of systems in Intelligent Computer-Assisted Language Learning (ICALL) is to provide immediate feedback to language learners working on exercises. Most of this research has focused on providing feedback on the form of the learner input. Foreign language practice and second language acquisition research, on the other hand, emphasizes the importance of exercises that require the learner to manipulate meaning.

The ability of an ICALL system to diagnose and provide feedback on the meaning conveyed by a learner response depends on how well it can deal with the response variation allowed by an activity. We focus on short-answer reading comprehension questions which have a clearly defined target response but the learner may convey the meaning of the target in multiple ways. As empirical basis of our work, we collected an English as a Second Language (ESL) learner corpus of short-answer reading comprehension questions, for which two graders provided target answers and correctness judgments. On this basis, we developed a Content-Assessment Module (CAM), which performs shallow semantic analysis to diagnose meaning errors. It reaches an accuracy of 88% for semantic error detection and 87% on semantic error diagnosis on a held-out test data set.

1 Introduction

Language practice that includes meaningful interaction is a critical component of many current language teaching theories. At the same time, existing research on intelligent computer-aided language learning (ICALL) systems has focused primarily on providing practice with grammatical forms. For most ICALL systems, although form assessment often involves the use of natural language processing (NLP) techniques, the need for sophisticated content assessment of a learner response is limited by restricting the kinds of activities offered in order to tightly control the variation allowed in learner responses, i.e., only one or very few forms can be used by the learner to express the correct content. Yet many of the activities that language instructors typically use in real language-learning settings support a significant degree of variation in correct answers and in turn require both form and content assessment for answer evaluation. Thus, there is a real need for ICALL systems that provide accurate content assessment.

While some meaningful activities are too unrestricted for ICALL systems to provide effective content assessment, where the line should be drawn on a spectrum of language exercises is an open question. Different language-learning exercises carry different expectations with respect to the level and type of linguistic variation possible across learner responses. In turn, these expectations may be linked to the learning goals underlying the activity design, the cognitive skills required to respond to the activity, or other properties of the activity. To develop adequate processing strategies for content assessment, it is important to understand the connection between exercises and expected variation, as conceptualized by the exercise spectrum shown in Figure 1, because the level of variation imposes re-

Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, pages 107–115,
Columbus, Ohio, USA, June 2008.

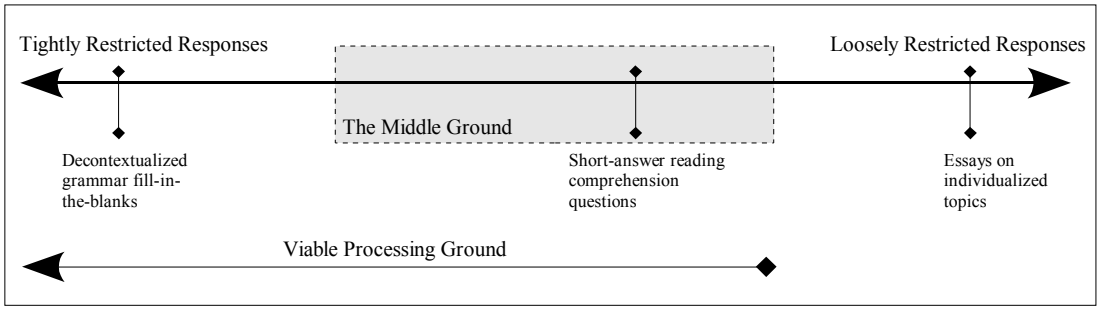

Figure 1: Language Learning Exercise Spectrum

quirements and limitations on different processing strategies. At one extreme of the spectrum, there are tightly restricted exercises requiring minimal analysis in order to assess content. At the other extreme are unrestricted exercises requiring extensive form and content analysis to assess content. In this work, we focus on determining whether shallow content-analysis techniques can be used to perform content assessment for activities in the space between the extremes. A good test case in this middle ground are loosely restricted reading comprehension (RC) questions. From a teaching perspective, they are a task that is common in real-life learning situations, they combine elements of comprehension and production, and they are a meaningful activity suited to an ICALL setting. From a processing perspective, responses exhibit linguistic variation on lexical, morphological, syntactic and semantic levels – yet the intended contents of the answer is predictable so that an instructor can define target responses.

Since variation is possible across learner responses in activities in the middle ground of the spectrum, we propose a shallow content assessment approach which supports the comparison of target and learner responses on several levels including token, chunk and relation. We present an architecture for a content assessment module (CAM) which provides this flexibility using multiple surface-based matching strategies and existing language processing tools. For an empirical evaluation, we collected a corpus of language learner data consisting exclusively of responses to short-answer reading comprehension questions by intermediate English language learners.

2 The Data

The learner corpus consists of 566 responses to short-answer comprehension questions. The responses, written by intermediate ESL students as part of their regular homework assignments, were typically 1-3 sentences in length. Students had access to their textbooks for all activities. For development and testing, the corpus was divided into two sets. The development set contains 311 responses from 11 students answering 47 different questions; the test set contains 255 responses from 15 students to 28 questions. The development and test sets were collected in two different classes of the same intermediate reading/writing course.

Two graders annotated the learner answers with a binary code for semantic correctness and one of several diagnosis codes to be discussed below. Target responses (i.e., correct answers) and keywords from the target responses were also identified by the graders.[1] Because we focus on content assessment, learner responses containing grammatical errors were only marked as incorrect if the grammatical errors impacted the understanding of the meaning.

The graders did not agree on correctness judgments for 31 responses (12%) in the test set. These were eliminated from the test set in order to obtain a gold standard for evaluation.

The remaining responses in the development and test sets showed a range of variation for many of the prompts. As the following example from the corpus illustrates, even straightforward questions based on

[1] Keywords refer to terms in the target response essential to a correct answer.

an explicit short reading passage yield both linguistic and content variation:

CUE: *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.*

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.*

While the third answer was judged to be correct, the syntactic structures, word order, forms, and lexical items used (e.g., *famous person* vs. *famous people*) vary from the string provided as target. Of the learner responses in the corpus, only one was string identical with the teacher-provided target and nine were identical when treated as bags-of-words. In the test set, none of the learner responses was string or bag-of-word identical with the corresponding target sentence.

To classify the variation exhibited in learner responses, we developed an annotation scheme based on target modification, with the meaning error labels being adapted from those identified by James (1998) for grammatical mistakes. Target modification encodes how the learner response varies from the target, but makes the sometimes incorrect assumption that the learner is actually trying to "hit" the meaning of the target. The annotation scheme distinguishes *correct answers*, *omissions* (of relevant concepts), *overinclusions* (of incorrect concepts), *blends* (both omissions and overinclusions), and *non-answers*. These error types are exemplified below with examples from the corpus. In addition, the graders used the label *alternate answer* for responses that were correct given the question and reading passage, but that differed significantly in meaning from what was conveyed by the target answer.[2]

1. Necessary concepts left out of learner response.

   CUE: *Name the features that are used in the design of advertisements.*
   TARGET: *The features are eye contact, color, famous people, language and cultural references.*
   RESPONSE: *Eye contact, color*

2. Response with extraneous, incorrect concepts.

   CUE: *Which form of programming on TV shows that highest level of violence?*
   TARGET: *Cartoons show the most violent acts.*
   RESPONSE: *Television drama, children's programs and cartoons.*

3. An incorrect blend/substitution (correct concept missing, incorrect one present).

   CUE: *What is alliteration?*
   TARGET: *Alliteration is where sequential words begin with the same letter or sound.*
   RESPONSE: *The worlds are often chosen to make some pattern or play on works. Sequential works begins with the same letter or sound.*

4. Multiple incorrect concepts.

   CUE: *What was the major moral question raised by the Clinton incident?*[3]
   TARGET: *The moral question raised by the Clinton incident was whether a politician's personal life is relevant to their job performance.*
   RESPONSE: *The scandal was about the relationship between Clinton and Lewinsky.*

3 Method

The CAM design integrates multiple matching strategies at different levels of representation and various abstractions from the surface form to compare meanings across a range of response variations. The approach is related to the methods used in

[2] We use the term *concept* to refer to an entity or a relation between entities in a representation of the meaning of a sentence. Thus, a response generally contains multiple concepts.

[3] Note the incorrect presupposition in the cue provided by the instructor.

machine translation evaluation (e.g., Banerjee and Lavie, 2005; Lin and Och, 2004), paraphrase recognition (e.g., Brockett and Dolan, 2005; Hatzivassiloglou et al., 1999), and automatic grading (e.g., Leacock, 2004; Marín, 2004).

To illustrate the general idea, consider the example from our corpus in Figure 2.

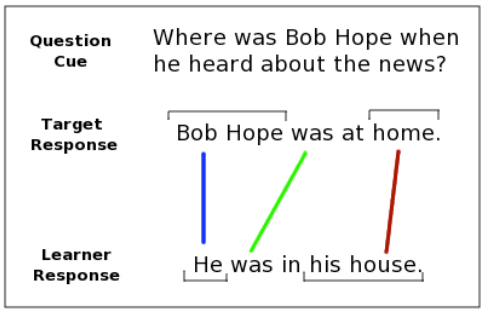

Figure 2: Basic matching example

We find one string identical match between the token *was* occurring in the target and the learner response. At the noun chunk level we can match *home* with *his house*. And finally, after pronoun resolution it is possible to match *Bob Hope* with *he*.

The overall architecture of CAM is shown in Figure 3. Generally speaking, CAM compares the learner response to a stored target response and decides whether the two responses are possibly different realizations of the same semantic content. The design relies on a series of increasingly complex comparison modules to "align" or match compatible concepts. Aligned and unaligned concepts are used to diagnose content errors. The CAM design supports the comparison of target and learner responses on token, chunk and relation levels. At the token level, the nature of the comparison includes abstractions of the string to its lemma (i.e., uninflected root form of a word), semantic type (e.g., date, location), synonyms, and a more general notion of similarity supporting comparison across part-of-speech.

The system takes as input the learner response and one or more target responses, along with the question and the source reading passage. The comparison of the target and learner input pair proceeds first with an analysis filter, which determines whether linguistic analysis is required for diagnosis. Essentially, this filter identifies learner responses that were copied directly from the source text.

Then, for any learner-target response pair that requires linguistic analysis, CAM assessment proceeds in three phases – Annotation, Alignment and Diagnosis. The Annotation phase uses NLP tools to enrich the learner and target responses, as well as the question text, with linguistic information, such as lemmas and part-of-speech tags. The question text is used for pronoun resolution and to eliminate concepts that are "given" (cf. Halliday, 1967, p. 204 and many others since). Here "given" information refers to concepts from the question text that are reused in the learner response. They may be necessary for forming complete sentences, but contribute no new information. For example, if the question is *What is alliteration?* and the response is *Alliteration is the repetition of initial letters or sounds*, then the concept represented by the word *alliteration* is given and the rest is new. For CAM, responses are neither penalized nor rewarded for containing given information.

Table 1 contains an overview of the annotations and the resources, tools or algorithms used. The choice of the particular algorithm or implementation was primarily based on availability and performance on our development corpus – other implementations could generally be substituted without changing the overall approach.

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, 1997)
Lexical Relations	WordNet (Miller, 1995)
Similarity Scores	PMI-IR (Turney, 2001; Mihalcea et al., 2006)
Dependency Relations	Stanford Parser (Klein and Manning, 2003)

Table 1: NLP Tools used in CAM

After the Annotation phase, Alignment maps new (i.e., not given) concepts in the learner response to concepts in the target response using the annotated information. The final Diagnosis phase analyzes the alignment to determine whether the learner re-

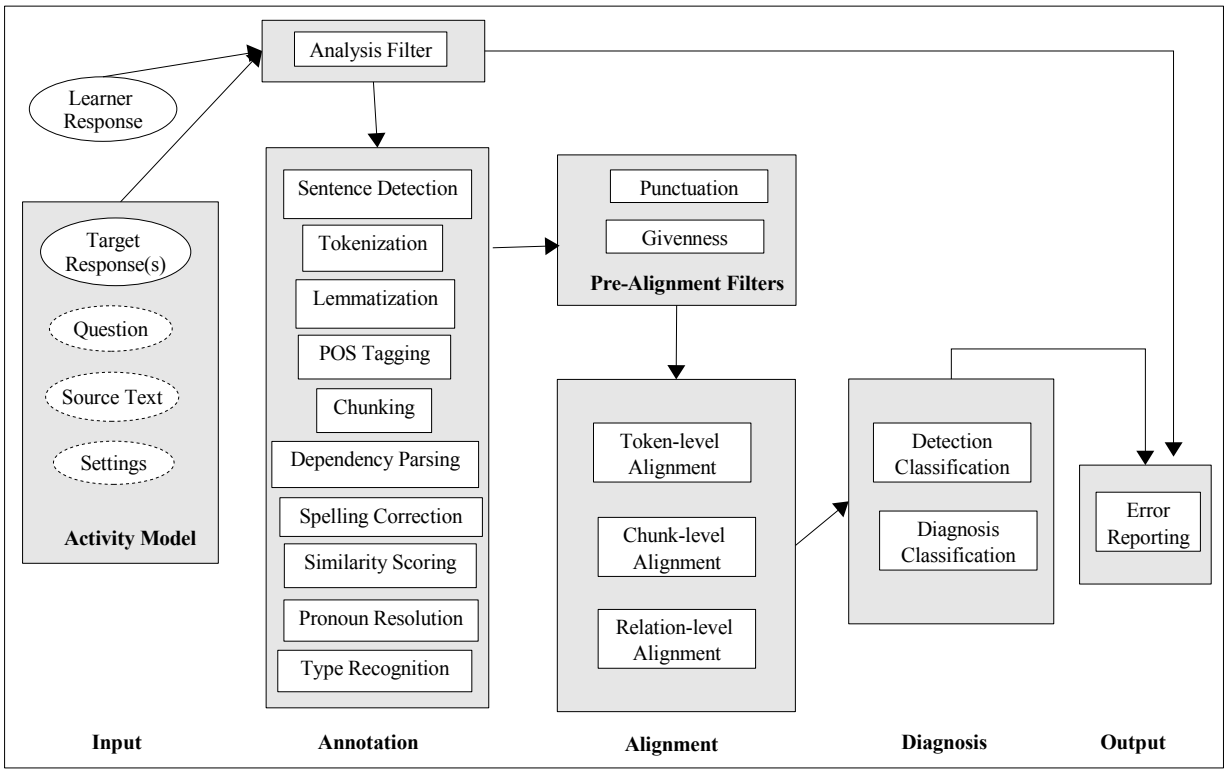

Figure 3: Architecture of the Content Assessment Module (CAM)

sponse contains content errors. If multiple target responses are supplied, then each is compared to the learner response and the target response with the most matches is selected as the model used in diagnosis. The output is a diagnosis of the input pair, which might be used in a number of ways to provide feedback to the learner.

3.1 Combining the evidence

To combine the evidence from these different levels of analysis for content evaluation and diagnosis, we tried two methods. In the first, we hand-wrote rules and set thresholds to maximize performance on the development set. On the development set, the hand-tuned method resulted in an accuracy of 81% for the semantic error detection task, a binary judgment task. However, performance on the test set (which was collected in a later quarter with a different instructor and different students) made clear that the rules and thresholds thus obtained were overly specific to the development set, as accuracy dropped down to 63% on the test set. The hand-written rules apparently were not general enough to transfer well from the development set to the test set, i.e., they relied on properties of the development set that where not shared across data sets. Given the variety of features and the many different options for combining and weighing them that might have been explored, we decided that rather than hand-tuning the rules to additional data, we would try to machine learn the best way of combining the evidence collected. We thus decided to explore machine learning, even though the set of development data for training clearly is very small.

Machine learning has been used for equivalence recognition in related fields. For instance, Hatzivassiloglou et al. (1999) trained a classifier for paraphrase detection, though their performance only reached roughly 37% recall and 61% precision. In a different approach, Finch et al. (2005) found that MT evaluation techniques combined with machine learning improves equivalence recognition. They used the output of several MT evaluation approaches based on matching concepts (e.g., BLEU) as features/values for training a support vector machine (SVM) classifier. Matched concepts and unmatched

concepts alike were used as features for training the classifier. Tested against the Microsoft Research Paraphrase (MSRP) Corpus, the SVM classifier obtained 75% accuracy on identifying paraphrases. But it does not appear that machine learning techniques have so far been applied to or even discussed in the context of language learner corpora, where the available data sets typically are very small.

To begin to address the application of machine learning to meaning error diagnosis, the alignment data computed by CAM was converted into features suitable for machine learning. For example, the first feature calculated is the relative overlap of aligned keywords from the target response. The full list of features are listed in Table 2.

Features	Description
1. Keyword Overlap	Percent of keywords aligned (relative to target)
2. Target Overlap	Percent of aligned target tokens
3. Learner Overlap	Percent of aligned learner tokens
4. T-Chunk	Percent of aligned target chunks
5. L-Chunk	Percent of aligned learner chunks
6. T-Triple	Percent of aligned target triples
7. L-Triple	Percent of aligned learner triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments that were similarity-resolved
10. Type Match	Percent of token alignments that were type-resolved
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments that were synonym-resolved
13. Variety of Match (0-5)	Number of kinds of token-level alignments

Table 2: Features used for Machine Learning

Features 1-7 reflect relative numbers of matches (relative to length of either the target or learner response). Features 2, 4, and 6 are related to the target response overlap. Features 3, 5, and 7 are related to overlap in the learner response. Features 8–13 reflect the nature of the matches.

The values for the 13 features in Table 2 were used to train the detection classifier. For diagnosis, a fourteenth feature – a detection feature (1 or 0 depending on whether the detection classifier detected an error) – was added to the development data to train the diagnosis classifier. Given that token-level alignments are used in identifying chunk- and triple-level alignments, that kinds of alignments are related to variety of matches, etc., there is clear redundancy and interdependence among features. But each feature adds some new information to the overall diagnosis picture.

The machine learning suite used in all the development and testing runs is TiMBL (Daelemans et al., 2007). As with the NLP tools used, TiMBL was chosen mainly to illustrate the approach. It was not evaluated against several learning algorithms to determine the best performing algorithm for the task, although this is certainly an avenue for future research. In fact, TiMBL itself offers several algorithms and options for training and testing. Experiments with these options on the development set included varying how similarity between instances was measured, how importance (i.e., weight) was assigned to features and how many neighbors (i.e., instances) were examined in classifying new instances. Given the very small development set available, making empirical tuning on the development set difficult, we decided to use the default learning algorithm (k-nearest neighbor) and majority voting based on the top-performing training runs for each available distance measure.

4 Results

Turning to the results obtained by the machine-learning based CAM, for the binary semantic error detection task, the system obtains an overall 87% accuracy on the development set (using the leave-one-out option of TiMBL to avoid training on the test item). Interestingly, even for this small development set, machine learning thus outperforms the accuracy obtained for the manual method of combining the evidence reported above. On the test set, the final TiMBL-based CAM performance for detection improved slightly to 88% accuracy. These results suggest that detection using the CAM design is viable, though more extensive testing with a larger corpus is needed.

Balanced sets Both the development and test sets contained a high proportion of correct answers – 71% of the development set and 84% of the test set were marked as correct by the human graders. Thus,

we also sampled a balanced set consisting of 50% correct and 50% incorrect answers by randomly including correct answers plus all the incorrect answers to obtain a set with 152 cases (development subset) and 72 (test subset) sentences. The accuracy obtained for this balanced set was 78% (leave-one-out-testing with development set) and 67% (test set). The fact that the results for the balanced development set using leave-one-out-testing are comparable to the general results shows that the machine learner was not biased towards the ratio of correct and incorrect responses, even though there is a clear drop from development to test set, possibly related to the small size of the data sets available for training and testing.

Alternate answers Another interesting aspect to discuss is the treatment of alternate answers. Recall that alternate answers are those learner responses that are correct but significantly dissimilar from the given target. Of the development set response pairs, 15 were labeled as alternate answers. One would expect that given that these responses violate the assumption that the learner is trying to hit the given target, using these items in training would negatively effect the results. This turns out to be the case; performance on the training set drops slightly when the alternate answer pairs are included. We thus did not include them in the development set used for training the classifier. In other words, the diagnosis classifier was trained to label the data with one of five codes – *correct*, *omissions* (of relevant concepts), *overinclusions* (of incorrect concepts), *blends* (both omissions and overinclusions), and *non-answers*. Because it cannot be determined beforehand which items in unseen data are alternate answer pairs, these pairs were not removed from the test set in the final evaluation. Were these items eliminated, the detection performance would improve slightly to 89%.

Form errors Interestingly, the form errors frequently occurring in the student utterances did not negatively impact the CAM results. On average, a learner response in the test set contained 2.7 form errors. Yet, 68% of correctly diagnosed sentences included at least one form error, but only 53% of incorrectly diagnosed ones did so. In other words, correct responses had more form errors than incorrect responses. Looking at numbers and combinations of form errors, no clear pattern emerges that would suggest that form errors are linked to meaning errors in a clear way. One conclusion to draw based on these data is that form and content assessment can be treated as distinct in the evaluation of learner responses. Even in the presence of a range of form-based errors, human graders can clearly extract the intended meaning to be able to evaluate semantic correctness. The CAM approach is similarly able to provide meaning evaluation in the presence of grammatical errors.

Diagnosis For diagnosis with five codes, CAM obtained overall 87% accuracy both on the development and on the test set. Given that the number of labels increases from 2 to 5, the slight drop in overall performance in diagnosis as compared to the detection of semantic errors (from 88% to 87%) is both unsurprising in the decline and encouraging in the smallness of the decline. However, given the sample size and few numbers of instances of any given error in the test (and development) set, additional quantitative analysis of the diagnosis results would not be particularly meaningful.

5 Related Work

The need for semantic error diagnosis in previous CALL work has been limited by the narrow range of acceptable response variation in the supported language activity types. The few ICALL systems that have been successfully integrated into real-life language teaching, such as German Tutor (Heift, 2001) and BANZAI (Nagata, 2002), also tightly control expected response variation through deliberate exercise type choices that limit acceptable responses. Content assessment in the German Tutor is performed by string matching against the stored targets. Because of the tightly controlled exercise types and lack of variation in the expected input, the assumption that any variation in a learner response is due to form error, rather than legitimate variation, is a reasonable one. The recently developed TAGARELA system for learners of Portuguese (Amaral and Meurers, 2006; Amaral, 2007) lifts some of the restrictions on exercise types, while relying on shallow semantic processing. Using strategies inspired by our work, TAGARELA incorporates simple content assessment for evaluating

learner responses in short-answer questions.

ICALL system designs that do incorporate more sophisticated content assessment include FreeText (L'Haire and Faltin, 2003), the Military Language Tutor (MILT) Program (Kaplan et al., 1998), and Herr Kommissar (DeSmedt, 1995). These systems restrict both the exercise types *and* domains to make content assessment feasible using deeper semantic processing strategies.

Beyond the ICALL domain, work in automatic grading of short answers and essays has addressed whether the students answers convey the correct meaning, but these systems focus on largely scoring rather than diagnosis (e.g., E-rater, Burstein and Chodorow, 1999), do not specifically address language learning contexts and/or are designed to work specifically with longer texts (e.g., AutoTutor, Wiemer-Hastings et al., 1999). Thus, the extent to which ICALL systems can diagnose meaning errors in language learner responses has been far from clear.

As far as we are aware, no directly comparable systems performing content-assessment on related language learner data exist. The closest related system that does a similar kind of detection is the C-rater system (Leacock, 2004). That system obtains 85% accuracy. However, the test set and scoring system were different, and the system was applied to responses from native English speakers. In addition, their work focused on detection of errors rather than diagnosis. So, the results are not directly comparable. Nevertheless, the CAM detection results clearly are competitive.

6 Summary

After motivating the need for content assessment in ICALL, in this paper we have discussed an approach for content assessment of English language learner responses to short answer reading comprehension questions, which is worked out in detail in Bailey (2008). We discussed an architecture which relies on shallow processing strategies and achieves an accuracy approaching 90% for content error detection on a learner corpus we collected from learners completing the exercises assigned in a real-life ESL class. Even for the small data sets available in the area of language learning, it turns out that machine learning can be effective for combining the evidence from various shallow matching features. The good performance confirms the viability of using shallow NLP techniques for meaning error detection. By developing and testing this model, we hope to contribute to bridging the gap between what is practical and feasible from a processing perspective and what is desirable from the perspective of current theories of language instruction.

References

Steven Abney, 1997. Partial Parsing via Finite-State Cascades. *Natural Language Engineering*, 2(4):337–344. http://vinartus.net/spa/97a.pdf.

Luiz Amaral, 2007. Designing Intelligent Language Tutoring Systems: Integrating Natural Language Processing Technology into Foreign Language Teaching. Ph.D. thesis, The Ohio State University.

Luiz Amaral and Detmar Meurers, 2006. Where does ICALL Fit into Foreign Language Teaching? Presentation at the 23rd Annual Conference of the Computer Assisted Language Instruction Consortium (CALICO), May 19, 2006. University of Hawaii. http://purl.org/net/icall/handouts/calico06-amaral-meurers.pdf.

Evan L. Antworth, 1993. Glossing Text with the PC-KIMMO Morphological Parser. *Computers and the Humanities*, 26:475–484.

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

Stacey Bailey, 2008. Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English as a Second Language. Ph.D. thesis, The Ohio State University.

Satanjeev Banerjee and Alon 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 Machine Translation and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*. Ann Arbor, Michigan, pp. 65–72. http://aclweb.org/anthology/W05-0909.

Chris Brockett and William 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. http://aclweb.org/anthology/I05-5001.

Jill Burstein and Martin 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. http://aclweb.org/anthology/W99-0411.

Walter Daelemans, 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.

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

Andrew Finch, Young-Sook Hwang and Eiichiro Sumita, 2005. Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. pp. 17–24. http://aclweb.org/anthology/I05-5003.

Michael Halliday, 1967. Notes on Transitivity and Theme in English. Part 1 and 2. *Journal of Linguistics*, 3:37–81, 199–244.

Vasileios Hatzivassiloglou, Judith Klavans and Eleazar 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. http://aclweb.org/anthology/W99-0625.

Trude Heift, 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.

Carl James, 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. Longman Publishers.

Jonathan Kaplan, 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.

Dan Klein and Christopher 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. http://aclweb.org/anthology/P03-1054.

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

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

Sébastien L'Haire and Anne Vandeventer Faltin, 2003. Error Diagnosis in the FreeText Project. *CALICO Journal*, 20(3):481–495.

Chin-Yew Lin and Franz Josef 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. http://aclweb.org/anthology/P04-1077.

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

Diana Rosario Pérez Marín, 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.

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

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

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

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

Peter Turney, 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.

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