Exercise-driven selection of content matching methodologies for ICALL

Stacey Bailey
Department of Linguistics
The Ohio State University

September 6, 2006

* Based on joint work with Detmar Meurers.

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 and Rodgers 2001).

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

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.

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.

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 and Nicholson 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.

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.

Shifting the perspective of ICALL system design

- Fortunately, NLP technology has progressed to the point of having tools available for analysis beyond form processing.
- 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.

Relating language exercises and NLP

- The more variation possible in learner responses to a language 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 form and meaning analysis to assess meaning.

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 1: Guided fill-in-the-blank*

> *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.

*Activity from Azar (1999), a grammar textbook for learners of American English.

Exercise 2: Open-ended questions*

> *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.
- Such activities are better suited to in-class settings.

*Activity from Kirn and Hartmann (2002), a textbook for learners of English.

The middle ground

- The space between the opposite ends of the spectrum could be a good compromise between what is practical and what is needed in ICALL activities.
- The degree to which exercises in the middle ground can be easily, effectively and reliably processed with NLP technology is what we are exploring.

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

> 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.

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

Exemplifying the middle ground: Question answering

> 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.

* Activity from Seal (1997).

Exemplifying the middle ground: Information gap

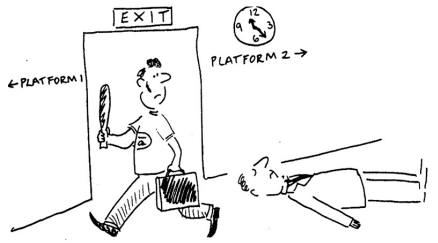
You will be asked questions…

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.

* Activity from Birch (2005).

Minimal NLP requirements

- Tokenization: from raw input to words.
- Morphological analysis: from words to stems/lemmas.
- Lexical resources: identifying word associations (synonyms, hyponyms, meronyms, etc.)
- Part of speech tagging: lexical category assignment.
- (Shallow) parsing: syntactic structure assignment.
- Shallow semantic analysis: identifying relations between concepts.

Reliability of NLP technology (1)

- The reliability of NLP components depends on factors such as the nature of the language, domain, task and specific implementation.
- These results are for models tested on English newspaper data.
  - Tokenization: 99.7% accuracy (Grefenstette and Tapanainen 1994)
    - Issues: *New York, Mass., four-dimensional, in spite of,* etc.
  - Part of Speech Tagging: 97% accuracy (Brants 1998)
    - Issues: *at* (preposition or particle?), *writing* (verb, adjective or noun?)
  - Parsing: 90+% accuracy
    - Issues: *An enraged cow killed a farmer with an axe.*
  - Named Entity Recognition: 93% (Mikheev, et al. 1999)
    - Issues: *Marx Brothers* (person or company?)
- NLP technology can be brittle when used on text of a different domain or ill-formed input.

Reliability of NLP technology (2)

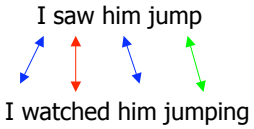
- The prospect of using imperfect technology is not necessarily grim:
  - Human performance on these tasks is often not 100%.
  - The types of errors each technology makes are not evenly distributed over all cases that technology must handle.
- Implications for ICALL system design:
  - Good activity design can help ICALL systems avoid those hard cases in which the technology is likely to fail.
  - Application of the most reliable technology first, whenever possible, can lessen the impact of unreliable technology.

A basic model for meaning assessment

- Our foundation for building a meaning assessment module is METEOR, a state-of-the-art system for machine translation (MT) evaluation (Banerjee and Lavie 2004).
- METEOR uses
  - a modularized structure for concept matching,
  - surface-level processing strategies, and
  - concept matching at the token, stem or synonym level.
- Given that these design features fulfill our criteria, we have implemented them in a basic model for meaning assessment.

Basic model processing example

- A sample target-response pair from a text corpus of Japanese learners of English (Miura 1998):
  - **Exercise**: A translation task from Japanese to English.
  - **Target Response**: *I saw him jump.*
  - **Learner Response**: *I watched him jumping*.
- Mapped concepts:



- The basic model selects token, stem and synonym alignments, in that order.

Basic model processing details

- The final alignment of concepts is used to determine the similarity between the target and learner responses.
- Any unaligned concepts in the learner and target responses can be evaluated to provide feedback for the meaning assessment.
- The assessment – how the aligned and unaligned elements are interpreted – is flexible, based on the goals of the activity.
  - For the translation evaluations, all the content words must be present and the structure of the learner response should be as close as possible to the target.

Building on the foundation

- We are extending the basic model to
  - Support alignments in the presence of a wider range of equivalence classes.
    - Phrasal verbs (*look at* vs. *watch*)
    - Morphological variation (*he* vs. *him*)
    - Multi-word tense expressions (*sat* vs. *was sitting*)
    - Etc.
  - Identify and align relations between concepts.
    - Arguments (Ex: *He sat watching the river* vs. *The river sat watching him.*)
    - Modifiers (Ex: *brown fox and lazy dog* vs. *lazy fox and brown dog*)
    - Coreference (Ex: *Mohandas Karamchand Gandhi, Gandhi, Mahatma Gandhi, Bapu*)
    - Etc.

Testing the model

- We are collecting actual learner responses to language activities in order to test the effectiveness of the model.
- The targeted language activities
  - are currently used at OSU as part of the ESL curriculum,
  - fall in the middle ground of the spectrum, and
  - reflect a range of exercise types so that we may evaluate the effectiveness of content processing for different ICALL activities.

Summary

- Meaning assessment is essential for better integration of ICALL systems.
- Existing ICALL systems emphasize form assessment, limiting their usefulness in real-life language teaching.
- To improve usefulness, ICALL systems must be able to process learner responses from less-restricted activities.
- Such activities fall in the middle ground of a spectrum of language activity types and the processing they require for meaning assessment.
- Defining this middle ground is a critical step in determining the feasibility of incorporating those activities into an ICALL system.
- To explore properties and processing requirements of activities in the middle ground, we are developing a meaning assessment system.
- This system builds on the machine translation evaluation system METEOR to allow for content assessment of a wide range of concepts and relations between concepts.

References

- Azar, Betty Schrampfer. 1999. *Understanding and Using English Grammar*, Third Edition. New York: Longman Publishers.
- Banerjee, Satanjeev, and Lavie, Alon. 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).
- Birch, Gregory. 2005. "Balancing fluency, accuracy and complexity." In Corony Edwards and Jane Willis (Eds.), *Teachers Exploring Tasks in English Language Teaching*. Palgrave Macmillan. pp. 228–239.
- Brants, Thorsten. 2000. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*, ANLP-2000, April 29 – May 3, 2000, Seattle, WA.
- Grefenstette, Gregory and Tapanainen, Pasi. 1994. In Proceedings of the 3rd International Conference on Computational Lexicography, Budapest. pp. 79—87.
- Heift, Trude and Nicholson, Devlan. 2001. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education* 12(4). pp. 310–325.
- Kirn, Elaine and Hartmann, Pamela. 2002. *Interactions 2: Reading*, Fourth Edition. New York: McGraw-Hill Contemporary.
- L'Haire, Sebastien, and Faltin, Anne Vandeventer. 2003. "Error diagnosis in the FreeText project." *CALICO Journal* 20(3): 481.
- McCarthy, Michael and O'Dell, Felicity. 1997. *Vocabulary in Use: Upper Intermediate*. New York: Cambridge University Press.
- Mikheev, Andrei, Moens, Marc and Grover, Claire. 1999. Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8.
- Miura, Shogo. 1998. "Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II).", 1998. http://home.hiroshima-u.ac.jp/d052121/eigo2.html. Last Modified 14 May, 1998.
- Nagata, Noriko. 2002. BANZAI: An Application of Natural Language Processing to Web Based Language Learning, CALICO Journal 19 (3), 583-599.
- Richards, Jack and Rodgers, Theodore. 2001. Approaches and Methods in Language Teaching, Second Edition. New York: Cambridge University Press.
- Seal, Bernard. 1997. *Academic Encounters, Reading, Study Skills and Writing: Human Behavior*. New York: Cambridge University Press.