Research Statement

Lee Becker

My research interests center around natural language processing, information retrieval, and machine learning, and I am motivated by the promise these areas hold for improving education, scientific research, and human computer interaction. Much of my work utilizes intelligent tutoring systems (Woolf, 2008) and adaptive educational technologies as a vehicle for improving the state-of-the-art in natural language processing and as a framework for applying data mining to investigate the mechanisms and factors that drive learning. I have also recently started investigating the use of natural language processing to extract information from patient data and clinical medical notes.

1 Intelligent Tutoring Systems and Dialogue Management

Many studies have shown the benefits of one-on-one tutoring in both human-to-human and human-to-computer contexts. Similarly, Intelligent Tutoring Systems (ITS) have proven effective for teaching a wide variety of subject matter, and studies have found ITS can improve student performance by as much as a letter grade. Advances in speech and language technologies have allowed ITS to offer a more intuitive and engaging environment for learning, which in turn enables students to form deeper connections to the material. Although the potential benefits are understood, there are still several technical barriers to making these systems effective and accessible to a wider audience.

Two major challenges for natural language based ITS are 1) making the dialogue natural and responsive and 2) reducing the amount of human effort associated with developing and deploying such systems. For the past four years, I have been investigating techniques to address these issues via the development of a Socratic-style ITS for elementary school science education called My Science Tutor (MyST) (Ward et al., 2011). To converse with a student, an ITS must first understand what he or she is saying. For MyST we address the natural language understanding problem by first modeling learning goals with a frame semantic representation that decomposes concepts into their constituent actions, events, and relationships. We then use semantic grammars to parse and recognize possible paraphrases in student speech. To create robust tutorial behavior we utilize proven techniques for dialogue management that have been adapted for the goal of eliciting meaningful student responses. In its final year of evaluation MyST has proven to be an effective educational supplement with students who used MyST showing significant improvement in learning gains over their peers who only received in-class instruction (Ward et al., 2012).

While MyST represents a step forward in natural language tutoring, its creation required significant human effort in authoring prompts and tuning dialogue behavior. Furthermore, its actions and prompts are lesson specific. My long-term research goal is to enable rapid creation of ITS through advancements in natural language processing (NLP) and machine learning. I envision future platforms that can automatically extract conceptual knowledge from raw text, learn tutoring behaviors directly from real-world examples, and refine pedagogical strategies based on assessment outcomes and user feedback. Such a framework would allow easy deployment of a tutor for any subject, while simultaneously allowing customization of tutor behavior for the learner's needs.

As a first step toward facilitating automatic induction of models of dialogue behavior that generalize across subject domains, I have developed the Dialogue Schema to Unify Speech and Semantics (DISCUSS) (Becker et al., 2011). This multilayered taxonomy is a rich intermediate representation that captures the semantics (meaning) and pragmatics (action) of dialogue turns. Traditionally most dialogue systems have relied on high-level dialogue acts, which are too coarse to accurately represent and characterize conversational intent. Alternatively, most semantics based representations look only at specific word-word relationships. The DISCUSS representation aims to bridge this knowledge gap by simultaneously accounting for the dialogue action, rhetorical form, and predicate semantics of a conversational utterance.

For my dissertation I am investigating how the DISCUSS representation can be used to extract more informative features for statistical machine learning algorithms. Specifically, I am developing models for two tasks 1) characterizing and assessing the quality of the interactions within a tutorial dialogue and 2) ranking potential follow-up questions to student responses. To assemble the data necessary for both tasks, I have trained two linguists to annotate dialogue utterances with DISCUSS tags, and I have managed them as they annotated several thousand utterances from a corpus of MyST dialogues. To gather gold-standard data suitable for training machine learning algorithms, I have also collected thousands of judgments of question and dialogue quality from several expert tutors.

In the dialogue characterization task my system analyzes an annotated dialogue and predicts ratings that correlate with expert tutor assessment. Instead of assigning a single score, my system can score a dialogue along 11 different dimensions including student engagement, tutor responsiveness, lesson coverage, and overall learning experience (Becker et al., (b)). For the second task, the goal is to rank the appropriateness of candidate questions within a specific point of a dialogue. As with the dialogue rating task, my system learns to rank questions from preference data collected from expert tutors. (Becker et al., submitted (c)). In both tasks, the systems produce ratings and rankings comparable to expert tutor assessment.

The key challenges in this work center on the uncertainty associated with subjective human judgments. Aside from measuring long term learning gains (a tenuous measure at best), there is no generally accepted approach for evaluating tutorial dialogue quality and success. Because I have no single source for ‘ground truth’, my research challenge resides in creating meaningful evaluations as much as they do in defining system behavior. Another complicating factor stems from the inherent noise associated with subjective measurements. Even a relatively simple task like labeling product reviews with positive or negative sentiment poses difficulty for human judges; with educational data notions of quality or goodness become even hazier. If not utilized correctly, this noisy data can easily confuse machine learning algorithms. I have addressed this in my own research by using knowledge representations that allow for better generalization. In the future I see potential in employing semi-supervised and active learning to better understand these issues of uncertainty.

From an educational perspective, my dialogue rating and question ranking systems can assist in a variety of applications such as identifying struggling students, discovering difficult concepts, assessing tutor performance, or training human tutors in a new pedagogy. From a machine-learning perspective, automatic computation of these metrics can be used to optimize and customize a system’s dialogue behavior for specific domains and tutorial styles.

These data and these systems open several potential research directions. Because collecting expert judgments and linguistic annotation is expensive, I am interested in exploring how judgment and annotation tasks can be simplified to allow for reliable data collection via crowdsourcing services like Amazon’s Mechanical Turk. Similarly, I see several opportunities to use actual student outcomes and real-time feedback to automatically induce more intelligent and more personalized tutoring behavior. I am also interested in investigating how we can use natural language technologies to organize and present the wide array of readily available material found online (Wikipedia, digital libraries, online lectures, etc.) in a fashion more conducive to learning.

2 Question Generation

My research in intelligent tutoring systems has created a natural bridge to an emerging area in NLP known as Question Generation (QG). Question generation is the process of automatically generating questions from various sources of knowledge including raw text, databases, and semantic representations. This technology has the potential to improve the user experience and drastically reduce the development effort associated with several applications including dialogue systems, educational testing and assessment, and assistive agents.

Since becoming a member of the QG community, I have conducted experiments that shed light on the critical issues related to testing and evaluation of question generation systems (Becker et al., 2009, 2010), and I subsequently helped to define a shared-task and evaluation challenge to help advance QG research and to promote the QG community to the broader NLP and ITS communities. My involvement in QG also led to an internship at Microsoft Research, where my collaborators and I developed a system that automatically generated quizzes from general texts like Wikipedia articles (Becker et al., (a)). Hoping to leverage the wisdom of the crowd, we approached this problem as a supervised machine learning task. To generate questions, we used a suite of NLP tools such as summarizers, syntactic parsers, semantic role labelers, and named entity recognizers to identify potentially important spans of texts for question generation. We then collected human judgments of question quality and relevance using Amazon's Mechanical Turk crowdsourcing service, which were then used to train machine learning algorithms to predict the utility of a question.

This system and other QG systems represent a building block for my overarching goal of developing adaptive learning systems to assist self-motivated learners in exploring, understanding and mastering concepts in a new domain. In the future, I would like to utilize QG-based learning systems as a platform for investigating the efficacy of different pedagogies for different learning styles, and as an instrument for creating data-driven models of student understanding. Furthermore, I see QG as a useful means to collect data for evaluating and improving research in other areas of NLP research including automatic summarization, paraphrase detection, and textual entailment recognition.

References

Becker, L., Basu, S., and Vanderwende, L. (2012a). Mind the Gap: Learning to Choose Gaps for Question Generation. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada.

Becker, L., Nielsen, R. D., Okoye, I., Sumner, T., and Ward, W. (2010). What's next? Target concept identification. In *Proceedings of the 3rd Workshop on Question Generation*, Pittsburgh, PA USA.

Becker, L., Nielsen, R. D., and Ward, W. H. (2009). What a pilot study says about running a question generation challenge. In *The 2nd Workshop on Question Generation*, Brighton, England.

Becker, L., Palmer, M., van Vuuren, S., and Ward, W. (2012b). Learning to tutor like a tutor: Ranking questions in context. In *The 11th International Conference on Intelligent Tutoring Systems*, Crete, Greece.

Becker, L., Palmer, M., van Vuuren, S., and Ward, W. (2012c). Question ranking and selection for tutorial dialogue. In *The 7th Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada.

Becker, L., Ward, W., van Vuuren, S., and Palmer, M. (2011). DISCUSS: A dialogue move taxonomy layered over semantic representations. In *IWCS 2011: The 9th International Conference on Computational Semantics*, Oxford, England.

Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., van Vuuren, S., Weston, T., Zheng, J., and Becker, L. (2011). My Science Tutor: A conversational multi-media virtual tutor for elementary school science. *ACM TSLP: Special Issue on Speech and Language Processing of Children's Speech for Child-machine Interaction Applications*, 7(1).

Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., van Vuuren, S., Weston, T., Zheng, J., and Becker, L. (2012). My Science Tutor: A conversational multimedia virtual tutor (submitted for review). *Journal of Educational Psychology*.

Woolf, B. (2008). *Building Intelligent Interactive Tutors*. Morgan Kaufman.