Chapter 6
The User-Language Paraphrase Corpus

Philip M. McCarthy
University of Memphis, USA

Danielle S. McNamara
University of Memphis, USA

ABSTRACT

The corpus in this challenge comprises 1998 target-sentence/student response text-pairs, or protocols. The protocols have been evaluated by extensively trained human raters; however, unlike established paraphrase corpora that evaluate paraphrases as either true or false, the User-Language Paraphrase Corpus evaluates protocols along 10 dimensions of paraphrase characteristics on a six point scale. Along with the protocols, the database comprising the challenge includes 10 computational indices that have been used to assess these protocols. The challenge posed for researchers is to describe and assess their own approach (computational or statistical) to evaluating, characterizing, and/or categorizing, any, some, or all of the paraphrase dimensions in this corpus. The purpose of establishing such evaluations of user-language paraphrases is so that ITSs may provide users with accurate assessment and subsequently facilitative feedback, such that the assessment would be comparable to one or more trained human raters. Thus, these evaluations will help to develop the field of natural language assessment and understanding (Rus, McCarthy, McNamara, & Graesser, 2008 [a]).

DOI: 10.4018/978-1-61350-447-5.ch006
THE NEED FOR ACCURATE USER-LANGUAGE EVALUATION

The User-Language Paraphrase Corpus is a freely available set of data designed to function as a challenge for researchers interested in creating or testing approaches to paraphrase evaluations. The authors use the term User-Language to refer to the natural language input of users interacting with an intelligent tutoring system (ITS). The primary characteristics of user-language are that the input is short (typically a single sentence) and that it is unedited (e.g., it is replete with typographical errors and lacking in grammaticality). The authors use the term paraphrase to refer to ITS users’ attempt to restate a given target sentence in their own words such that a produced sentence, or user response, has the same meaning as the target sentence.

Intelligent Tutoring Systems (ITSs) are automated tools that implement systematic techniques for promoting learning (e.g., Aleven & Koedinger, 2002; Gertner & VanLehn, 2000; McNamara, Levinstein, & Boonthum, 2004). A subset of ITSs also incorporate conversational dialogue components that rely on computational linguistic algorithms to interpret and respond to natural language input by the user (see Rus et al., 2008 [a]). The computational algorithms enable the system to track students’ performance and adaptively respond. As such, the accuracy of the ITS responses to the user critically depends on the system’s interpretation of the user-language (McCarthy et al., 2007; McCarthy et al., 2008; Rus et al., 2008 [a]).

ITSs often assess user-language via one of several systems of matching. For instance, the user input may be compared against a pre-selected stored answer to a question, solution to a problem, misconception, target sentence/text, or other form of benchmark response (McNamara et al., 2007; Millis et al. 2007). Examples of systems that incorporate these approaches include AutoTutor, Why-Atlas, and iSTART (Graesser, et al. 2005; McNamara, Levinstein, & Boonthum, 2004; VanLehn et al., 2007). Although systems such as these vary widely in their goals and composition, ultimately their feedback mechanisms depend on comparing one text against another and forming an evaluation of their degree of similarity.

THE SEVEN MAJOR PROBLEMS WITH EVALUATING USER-LANGUAGE

Although a wide variety of tools and approaches have assessed edited, polished texts with considerable success, research on the computational assessment of ITS user-language textual relatedness has been less common and is less developed. As ITSs become more common, the need for accurate, yet fast evaluation of user-language becomes more pressing. However, meeting this need is challenging. This challenge is due, at least partially, to seven characteristics of user-language that complicate its evaluation.

Text Length

User-language is often short, typically no longer than a sentence. Established textual relatedness indices such as latent semantic analysis (LSA; Landauer et al., 2007) operate most effectively over longer texts where issues of syntax and negation are able to wash out by virtue of an abundance of commonly co-occurring words. Over shorter lengths, such approaches tend to lose their accuracy, generally correlating with text length (Dennis, 2007; McCarthy et al., 2007; McNamara et al., 2006; Penumatsa et al., 2004; Rehder et al. 1998; Rus et al., 2007; Wiemer-Hastings, 1999). The result of this problem is that longer responses tend to be judged more favorably in an ITS environment. Consequently, a long (but wrong) response may receive more favorable feedback than one that is short (but correct).
Related Content

Recognizing Prosody from the Lips: Is It Possible to Extract Prosodic Focus from Lip Features?
www.igi-global.com/chapter/recognizing-prosody-lips/31076?camid=4v1a

Semantic Methods for Textual Entailment
Andrew Neel and Max H. Garzon (2012). Applied Natural Language Processing: Identification, Investigation and Resolution  (pp. 479-494).
www.igi-global.com/chapter/semantic-methods-textual-entailment/61066?camid=4v1a

From Existential Graphs to Conceptual Graphs
www.igi-global.com/chapter/from-existential-graphs-to-conceptual-graphs/108732?camid=4v1a

Itakura-Saito Nonnegative Factorizations of the Power Spectrogram for Music Signal Decomposition
www.igi-global.com/chapter/itakura-saito-nonnegative-factorizations-power/45489?camid=4v1a