Chapter 26
Maximizing ANLP Evaluation:
Harmonizing Flawed Input

Adam M. Renner
The University of Memphis, USA

Philip M. McCarthy
The University of Memphis, USA

Chutima Boonthum-Denecke
Hampton University, USA

Danielle S. McNamara
Arizona State University, USA

ABSTRACT

A continuing problem for ANLP (compared with NLP) is that language tends to be more natural in ANLP than that examined in more controlled natural language processing (NLP) studies. Specifically, ineffective or misleading feedback can result from faulty assessment of misspelled words. This chapter describes the Harmonizer system for addressing the problem of user input irregularities (e.g., typos). The Harmonizer is specifically designed for Intelligence Tutoring Systems (ITSs) that use NLP to provide assessment and feedback based on the typed input of the user. Our approach is to "harmonize" similar words to the same form in the benchmark, rather than correcting them to dictionary entries. This chapter describes the Harmonizer, and evaluates its performance using various computational approaches on unedited input from high school students in the context of an ITS (i.e., iSTART). Our results indicate that various metric approaches to NLP (such as word-overlap cohesion scores) are moderately affected when student errors are filtered by the Harmonizer. Given the prevalence of typing errors in the sample, the study substantiates the need to "clean" typed input in comparable NLP-based learning systems. The Harmonizer provides such ability and is easy to implement with light processing requirements.

DOI: 10.4018/978-1-60960-741-8.ch026

INTRODUCTION

Technologies designed to provide interactive learning environments are a vital and growing aspect of research and development in Applied Natural Language Processing (ANLP). Computerized learning systems, in particular Intelligent Tutoring Systems (ITSs), offer the potential to substantially impact learning through student modeling, adaptive feedback, interactivity, and engaging learning environments. One means of providing adaptive tutoring within ITSs is through conversational dialogue between the computer interface and the student. In order for an ITS to successfully engage the student in meaningful conversational interactions, NLP algorithms can be used to evaluate students' natural language input into the system and direct responses to the student. In such systems, ITSs must rely on statistical and algebraic representations of human language to determine the most appropriate feedback to provide to the student. Successful and appropriate interactions between the student and an ITS are those in which the ITS accurately ascertains what the student *intended*. Such interactions are assumed to enhance both learning and motivation on the part of the student (Koedinger & Anderson, 1997). However, students do not always type or say what they mean. This aspect of natural dialogue renders accurate interpretations challenging because NLP techniques are not always designed to be used in naturalistic applications. As such, one major problem of ANLP within computerized educational programs such as ITSs is that many algorithms' accuracy can be compromised during interactions with real students, particularly less skilled students. As research progresses and more *intelligent* ITSs are developed to include increasingly sophisticated interactivity and adaptability, they must also utilize NLP techniques that are accurate and efficient, and thus applicable to students of all proficiencies.

The focus of this chapter is the optimization of a selection of established NLP techniques that have been applied within ITS environments. The growth of research in computational linguistics has led to major advances in development of NLP indices for evaluating edited, publishable texts (Foltz, Gilliam, & Kendall, 2000; Foltz & Wells, 1999). Although NLP techniques have been well-established for assessing *clean* texts, they have been less prevalent and less well developed for assessing *user-language* (i.e., typed input during interactions with an ITS; McCarthy & McNamara, 20011). This lack of progress is due, at least partially, to characteristics of user-language that complicate its evaluation. Consequently, the application of many NLP techniques (e.g., LSA, Entailer) may be less appropriate for assessing student language, which is often riddled with typographical and grammatical mistakes (McNamara, Boonthum, Levinstein, & Millis, 2007).

BACKGROUND

ITSs often assess user-language via *matching* principles. For instance, user input is compared to a pre-selected benchmark response (e.g., *ideal answer*, *solution to a problem*, *misconception*, *target sentence/text*) by measuring content word overlap or semantic similarity (McNamara et al., 2007). Systems that use this principle include AutoTutor (Graesser et al., 1999), Why2-Atlas (VanLehn et al., 2007), and iSTART (McNamara, Levenstein, & Boonthum, 2004). Although ITSs vary widely in their goals and composition, ultimately their feedback systems rely upon comparing one text against another and evaluating their degree of similarity. Similarity assessments may falter when dealing with user-language, which is usually unedited and abundant with typographical errors and poor grammar. For instance, a word in a target sentence that a user intended to type may not be