Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 2012, Vol. 12, No. 2, pp. 38-40
The Forum

The Application of Natural Language Processing and Automated Scoring in Second Language Assessment

Heidi Han-Ting Liu

Teachers College, Columbia University

Natural language processing (NLP) is an area of research that is used to investigate the application of natural language and is the foundation of machine translation, natural language text processing, natural language generation, multilingual and cross language information retrieval, speech recognition, parsing, and expert systems. To understand natural language in order to build or select appropriate algorithms for processing, three major issues are called into attention: humans' thought processes, the meaning of linguistic input in context, and world knowledge. These considerations have led to the development of various types of NLP tools for lexical and morphological analysis, semantic and discourse analysis, as well as knowledge-based approaches (c.f., Chowdhury, 2003). After decades of evolution and advancement, the current stage of NLP, as Xi (2010) pointed out, has allowed language testing researchers to apply its techniques in developing automated scoring systems for the purpose of language learning and assessment.

The algorithms of NLP provide automated scoring systems a solid theoretical ground. Automated scoring systems have been adopted mainly for two kinds of language assessments: writing (i.e., essay scoring) and speaking (i.e., speech scoring). Automated essay scoring systems are generally designed to identify examinees' written production features in terms of fluency (the number of words in the essay), diction (the variation in word length), and syntactic complexity (the number of various parts of speech). Several expert essay-scoring systems have been published, such as PEG (Project Essay Grade), IEA (Intelligent Essay Assessor), BETSY (Bayesian Essay Test Scoring sYstem), and IntelliMetric. Among all, the most well-known system is perhaps e-rater, developed by Educational Testing Service (ETS). The original intent of e-rater was for it to serve as a second rater in the Analytic Writing Assessment (AWA) in GMAT; currently, e-rater is also used as a second rater in the analytic writing section in GRE, the independent writing task in TOEFL iBT, and as the sole rater for TOEFL online practice tests. Interestingly, there have not been as many expert systems for automated speech scoring compared to those for automated essay scoring. To date, only two have been more commonly applied in language assessment: Versant Tests by Ordinate Corporation and SpeechRater, by Educational Testing Service (ETS). Versant Tests aim to assess examinees' everyday listening and speaking ability by computing their test scores in listening vocabulary, repeat accuracy, reciting and pronunciation, reading fluency, and repeat fluency. The scoring system also takes suprasegmental features (e.g., timing, pause, rhythm, etc) into account. According to Ordinate Corporation's internal research, Versant Tests allow for high level of test administration efficiency, and their test results yield high reliability as well as high predictability of examinees' performance in real life (Townshend & Todic, 1999). However, other researchers (c.f., Bernstein, 1999; Xi, Higgins, Zechner, & Williamson, 2008) have pointed out that the task types adopted in Versant Tests limit the representativeness of communicative competence since a lot of the higher-order cognitive abilities or complex linguistic knowledge are not present. SpeechRater, on the other hand, was developed by ETS specifically for scoring the speaking

section in TOEFL iBT. The system collects potentially meaningful features of speech input for each test task to build a scoring model, which include length of silences per word, speaking rate in words per sound, average chunk length in words, etc (Xi et al., 2008). The scores reflect examinees' communicative competence in an academic setting.

How have automated scoring systems become so immensely involved in the field of second language assessment? The history can be traced back to the 1960s, when Page proposed to the College Entrance Examination Board (CEEB) that automated essay graders should be developed to share the heavy workload of English teachers. The design and development of automated scoring systems is basically efficiency- and usefulness- oriented. Research on the incorporation of NLP in automated scoring systems has demonstrated that it is a complicated and difficult task for the machines to identify learner language. Most of the automated scoring systems currently in use have established satisfying reliability with the human raters. For example, e-rater can reach an 87% to 97% agreement with human raters, and the exact and adjacent agreement rates between SpeechRater and human raters can be up to 99%.

However, what most researchers are more concerned about is the validity issue in automated scoring. Xi (2010) drew out ten fundamental questions that help detect the potential validity violation when using automated scoring systems, such as whether "the use of assessment tasks constrained by automated scoring technologies lead to construct under- or misrepresentation" or if "automated scoring yield scores that are sufficiently consistent across measurement contexts" (p.293). Weigle (2010) also reminded us that, so far, the NLP technologies are not mature enough to allow automated scoring systems to score organization, coherence, content and meaning, as well as cultural awareness the same way as human raters. Therefore, cautious steps still need to be taken when adopting automatic scoring systems in second language assessment.

With the aid of NLP, automatic scoring systems have provided the advantages of efficiency in test administration as well as immediate availability of results and feedback for test users and language learners. However, we should bear in mind that languages are complicated, let alone in a second language setting. At the current stage, it is not yet possible for NLP techniques to cleanly characterize well-formed utterances from ill-formed ones, and many learner language (or rather, interlanguage) features cannot be captured, which leads to most practitioners' belief that the existence of human raters is still necessary. In the future, besides perfecting the linguistic model used in automatic scoring and addressing the validity issues, ongoing research should also investigate the stability of the results and feedback provided by automatic scoring systems across various performance samples from learners of different backgrounds, as well as the effects of automatic scoring systems on second language learning.

REFERENCES

Bernstein, J. (1999). *PhonePassTM testing: Structure and construct.* Menlo Park, CA: Ordinate.

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27,* 301-315.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology, 37,* 51-89.

Townshend, B., & Todic, O. (1999). *Comparison of PhonePassTM Testing with the Educational Testing Service Test® of Spoken EnglishTM (TSE®)*. Menlo Park, CA: Ordinate.

Weigle, S. C. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing. *Language Testing, 27,* 335–353.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*, 291-300.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M.(2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (ETS Research Rep. No. RR-08-62). Princeton, NJ: ETS.