Computers in Human Behavior xxx (2017) 1–11

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Comparative evaluation of automated scoring of syntactic competence of non-native speakers

Klaus Zechner [a], Su-Youn Yoon [a], Suma Bhat [b], Chee Wee Leong [a, *]

[a] NLP & Speech Group, Educational Testing Service, Princeton, NJ 08541, USA
[b] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL 61820, USA

ARTICLE INFO

Article history:
Received 5 July 2016
Received in revised form
16 January 2017
Accepted 29 January 2017
Available online xxx

Keywords:
Automated scoring
Automated speech recognition
English language assessment

ABSTRACT

Syntactic competence, especially the ability to use a wide range of sophisticated grammatical expressions, represents an important aspect of communicative acumen. This paper explores the question of how to best evaluate the syntactic competence of non-native speakers in an automated way. Using spoken responses of test takers participating in an English practice assessment, three classes of grammatical features – features based on n-grams of part-of-speech tags (POS), features based on various clause types, and features based on various phrases – are compared in an end-to-end assessment system. Feature correlations with human proficiency scores show that POS features and phrase features exhibit the highest correlations with human scores. Including these three classes of grammar features in a baseline scoring model that measures various aspects of spoken proficiency excluding aspects of grammar, we find substantial increases in agreement between machine and human scores. Finally, we discuss the broader implications of our results on the design of automatic scoring systems for spoken language.

1. Introduction

English is becoming an international lingua franca for academia, business, tourism, and trade. As a result, the need for learning English as a second language (ESL), as well as for assessing English proficiency among learners, has increased in recent decades. The most common approach to language assessment has been to evaluate the four essential language modalities of a non-native speaker, namely his or her reading, listening, writing, and speaking abilities.[1]

The assessment of reading and listening skills is done indirectly, most often by means of multiple-choice questions on the content of a text or listening passage. A meaningful assessment of the two productive modalities, writing and speaking, is more challenging since they do not lend themselves easily to a multiple-choice testing paradigm. Rather, the test taker is typically asked to generate a so-called "constructed response" (CR) to a prompt containing a combination of various written, spoken, and/or visual stimuli, such as a lecture or a narrative text.

These highly varied CRs pose a fundamental challenge to assessment reliability in addition to raising concerns about consistency, validity, and fairness. A traditional way of carrying out scoring of this kind of assessment has been to recruit pools of trained human raters to apply 'rubrics,' which describe the typical characteristics of CRs at different score levels. One example of this type of instrument is the Rubrics for TOEFL Speaking.[2] Using human raters for CR scoring, however, has a number of disadvantages (Engelhard, 1994, 2002, pp. 261–287) including rater inconsistency, rater drift, central tendency, raters being too lenient or harsh, time needed to do the rating, rating cost, and complexity of rater scheduling. For these reasons, it has long been a goal to automate the scoring process of CRs. It should be mentioned, though, that automated scoring has its own challenges, as well. It is a difficult task to achieve a comprehensive analysis of all aspects of language proficiency that are deemed important for assessing a test taker. While some aspects (e.g., the rate of speech, the fluency of a spoken

* Corresponding author.
E-mail address: cleong@ets.org (C.W. Leong).

[1] Educational testing Service (2016). Test of English as a Foreign Language (TOEFL). https://www.ets.org/toefl/, Test of English for International Communication (TOEIC). https://www.ets.org/toeic/, Pearson (2016). Pearson Test of English - academic. http://www.pearsonpte.com/.

[2] Educational testing Service (2016). Rubrics for TOEFL Speaking. https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf.

http://dx.doi.org/10.1016/j.chb.2017.01.060
0747-5632/

response) may be computed quite straightforwardly, other aspects (e.g., the logic of an argument in an essay) may be much harder to evaluate with automated means. Work on automated scoring of written responses dates back more than 50 years (Page, 1966), and has led to the operational use of automated scoring (in conjunction with human raters) to evaluate written responses in the context of a high-stakes, international assessment of academic English for non-native writers (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012).

As for scoring spoken responses automatically, research has trailed significantly behind, beginning only around 1990, when the technology for automatic speech recognition (ASR) reached a level of performance that made such an undertaking feasible. Initial systems focused on very narrow definitions of spoken proficiency, such as reading a passage aloud or orally repeating an acoustic stimulus (Cucchiarini, Strik, & Boves, 1997, 1998; Bernstein, Cohen, Murveit, Rtischev, & Weintraub, 1990; Franco, Neumeyer, Kim, & Ronen, 1997). Even today, systems that handle these restricted forms of speech are the most widely used systems for scoring spoken responses. Technology for scoring spontaneous speech, on the other hand, is substantially more complex to create. One major challenge for the development of such systems is the availability of highly accurate ASR systems for spontaneous speech.

While automated assessment of restricted, predictable spoken responses can be performed with a high degree of accuracy (Balogh et al., 2012), the automated scoring of spontaneous and more open-ended speech still poses significant challenges. This task requires the evaluation of a much larger set of components of spoken proficiency. An example of such a component for spontaneous speech would be the diversity and accuracy of production of grammatical structures, such as sentences or phrases. This ability is considered to be a significant marker of second language learners' spoken proficiency.

Studies in automated speech scoring have focused on the measurement of several dimensions of speech production, including fluency (Cucchiarini, Strik, & Boves, 2000, 2002), pronunciation (Franco et al., 1997; Neumeyer, Franco, Digalakis, & Weintraub, 2000; Witt & Young, 1998; Witt, 1999), and prosody (Chen & Zechner, 2011). Though the influence of syntactic competence on second language proficiency in the context of manual assessment of oral responses (Halleck, 1995; Iwashita, Brown, McNamara, & O'Hagan, 2008; Iwashita, Prior, Watanabe, & Lee, 2010) is well understood, studies in the area of automated speech scoring have only recently begun to actively investigate measurement of grammar usage in spontaneous non-native speech (Bernstein, Cheng, & Suzuki, 2010; Bhat & Yoon, 2015; Chen, Tetreault, & Xi, 2010; Chen & Zechner, 2011). These recent studies have suggested different approaches for computing features from spoken responses, measuring various aspects of grammatical competence. The current work addresses a need that has been critically lacking — a comparative evaluation of the merits of these different approaches for the automated assessment of syntactic competence, especially the range and sophistication in the grammar usage, in non-native spontaneous speech. It investigates the extent to which several types of grammar features can be combined to increase scoring accuracy for spoken responses.

This paper is organized as follows. In Section 2, a summary of relevant prior work is provided, motivating the four research questions that we are addressing in our paper; in Section 3, we present the data and method used for our study, including a detailed description of all grammatical features computed on spoken responses; Section 4 provides the results of our study, and a detailed discussion of the findings is presented in Section 5.

2. Syntactic complexity and proficiency

In the domain of second language acquisition, "the *range* of forms that surface in language production and the degree of *sophistication* of such forms" are considered to be two important aspects of grammar usage, collectively termed "syntactic complexity" (Ortega, 2003). Several measures of syntactic complexity have been used as indicators of a speaker's level of acquisition of syntactic competence for manual evaluations. In turn, such measures are suggestive of proficiency levels in second language writing and speaking (Halleck, 1995; Iwashita et al., 2010, 2008; Lu, 2010; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998).

These measures can be broadly classified into two groups (Bardovi-Harlig & Bofman, 1989):

- The *expression-based* group is focused on the frequency of specific, well-formed grammatical structures, such as negation structures and relative clauses. These measures are concerned with the acquisition of specific morpho-syntactic features or grammatical expressions characteristic of language acquisition stages.
- The *length-based* group, which is not restricted to particular structures, is related to the length of clauses or the relationship between clauses. Representative measures in this group include the *mean length of clause unit*, the *ratio of dependent clauses to the total number of clauses*, and the *number of verb phrases per clause.*

In contrast to syntactic complexity, grammatical accuracy is the ability to generate sentences without grammatical errors. The measures in this group can again be classified into two groups:

- *Global accuracy measures*, which include those that count all errors in sentence production and are calculated as normalized values, e.g., the percentage of error-free clauses among all clauses (Foster & Skehan, 1996).
- *Construction-specific accuracy measures*, which are focused on specific types of constructions such as verb tense, third-person singular forms, prepositions, and articles. The values of this measure are calculated as the percentage of error-free clauses with respect to these constructions (Iwashita et al., 2008; Robinson, 1995).

With respect to analyzing non-native spontaneous speech, instances of the above measures that are difficult to analyze, such as phrases without subjects or verbs, are plentiful. Foster, Tonkyn, and Wigglesworth (2000) provides many examples showing that a consistent application of the aforementioned measures to non-native speech is not easy to accomplish. A second difficulty is posed by the length of spoken responses, which are typically shorter than written responses. As a result, most measures based on sentence or sentence-like units are less reliable for use in speaking tasks that elicit only a few sentences. Not surprisingly, a marked decrease in the correlation between measures of syntactic complexity and proficiency as response length decreased was reported by Chen and Yoon (2011).

These issues are only compounded by related practical difficulties encountered when processing non-native spontaneous speech automatically, without any human annotations. Most measures used in related prior studies were based on production units, such as clauses, that were all manually identified. The task of automatically identifying them in speech that is naturally marked by frequent occurrences of fragments and ellipses render the measurement process difficult. Additionally, speech from English

Please cite this article in press as: Zechner, K., et al., Comparative evaluation of automated scoring of syntactic competence of non-native speakers, Computers in Human Behavior (2017), http://dx.doi.org/10.1016/j.chb.2017.01.060