Document Similarity Measures to Distinguish Native vs. Non-Native Essay Writers

Olga Gurevich
Educational Testing Service
Rosedale & Carter Roads,
Turnbull 11R
Princeton, NJ 08541
ogurevich@ets.org

Paul Deane
Educational Testing Service
Rosedale & Carter Roads,
Turnbull 11R
Princeton, NJ 08541
pdeane@ets.org

Abstract

The ability to distinguish statistically different populations of speakers or writers can be an important asset in many NLP applications. In this paper, we describe a method of using document similarity measures to describe differences in behavior between native and non-native speakers of English in a writing task.[1]

1 Introduction

The ability to distinguish statistically different populations of speakers or writers can be an important asset in many NLP applications. In this paper, we describe a method of using document similarity measures to describe differences in behavior between native and non-native speakers of English in a prompt response task.

We analyzed results from the new TOEFL integrated writing task, described in the next section. All task participants received the same set of prompts and were asked to summarize them. The resulting essays are all trying to express the same "gist" content, so that any measurable differences between them must be due to differences in individual language ability and style. Thus the task is uniquely suited to measuring differences in linguistic behavior between populations.

Our measure of document similarity, described in section 3, is a combination of word overlap and syntactic similarity, also serving as a measure of syntactic variability. The results demonstrate significant differences between native and non-native speakers that cannot be attributed to any demographic factor other than the language ability itself.

2 TOEFL Integrated Writing Task and Scoring

The Test of English as a Foreign Language (TOEFL) is administered to foreign students wishing to enroll in US or Canadian universities. It aims to measure the extent to which a student has acquired English; thus native speakers should on average perform better on the test regardless of their analytical abilities. The TOEFL now includes a writing component, and pilot studies were conducted with native as well as non-native speakers.

One of the writing components is an Integrated Writing Task. Students first read an expository passage, which remains on the screen throughout the task. Students then hear a segment of a lecture concerning the same topic. However, the lecture contradicts and complements the information contained in the reading. The lecture is heard once; students then summarize the lecture and the reading and describe any contradictions between them.

The resulting essays are scored by human raters on a scale of 0 to 5, with 5 being the best possible score[2]. The highest-scoring essays express ideas from both the lecture and the reading using correct grammar; the lowest-scoring essays rely on only one of the prompts for information and have grammatical problems; and the scores in between show a combination of both types of deficiencies.

The test prompt contained passages about the advantages and disadvantages of working in groups; the reading was 260 words long, the lecture 326 words. 540 non-native speakers and 950

[1] This research was funded while the first author was a Research Postdoctoral Fellow at ETS in Princeton, NJ.

[2] Native speaker essays were initially scored with possible half-grades such as 2.5. For purposes of comparison, these were rounded down to the nearest integer.

Proceedings of NAACL HLT 2007, Companion Volume, pages 49–52,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics

native speakers were tested by ETS in 2004. ETS also collected essential demographic data such as native language, educational level, etc., for each student. For later validation, we excluded 1/3 of each set, selected at random, thus involving 363 non-native speakers and 600 native speakers.

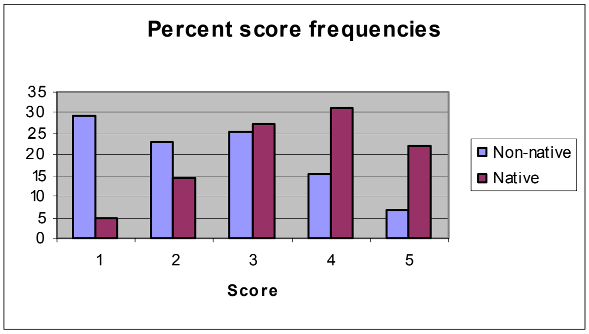

Figure 1. Relative score distributions.

Among the non-native speakers, the most common score was 1 (see Fig. 1 for a histogram). By contrast, native speaker scores centered around 3 and showed a normal-type distribution. The difference in distributions confirms that the task is effective at separating non-native speakers by skill level, and is easier for native speakers. The potential sources of difficulty include comprehension of the reading passage, listening ability and memory for the lecture, and the analytical ability to find commonalities and differences between the content of the reading and the lecture.

3 Document Similarity Measure

Due to the design of the TOEFL task, the content of the student essays is highly constrained. The aim of the computational measures is to extract grammatical and stylistic differences between different essays. We do this by comparing the essays to the reading and lecture prompts. Our end goal is to determine to what extent speakers diverge from the prompts while retaining the content.

The prediction is that native speakers are much more likely to paraphrase the prompts while keeping the same gist, whereas non-native speakers are likely to either repeat the prompts close to verbatim, or diverge from them in ways that do not preserve the gist. This intuition conforms to previous studies of native vs. non-native speakers' text summarization (cf. Campbell 1987), although we are not aware of any related computational work.

We begin by measuring lexico-grammatical similarity between each essay and the two prompts. An essay is represented as a set of features derived from its lexico-grammatical content, as described below. The resulting comparison measure goes beyond simple word or n-gram overlap by providing a measure of structural similarity as well. In essence, our method measures to what extent the essay expresses the content of the prompt in the same words, used in the same syntactic positions.

3.1 C-rater tuples

In order to get a measure of syntactic similarity, we relied on C-rater (Leacock & Chodorow 2003), an automatic scoring engine developed at ETS. C-rater includes several basic NLP components, including POS tagging, morphological processing, anaphora resolution, and shallow parsing. The parsing produces *tuples* for each clause, which describe each verb and its syntactic arguments (1).

(1) CLAUSE: the group spreads responsibility for a decision to all the members
TUPLE: *:verb:* spread *:subj:* the group *:obj:* responsible *:pp.for:* for a decide *:pp.to:* to all

C-rater does not produce full-sentence trees or prepositional phrase attachment. However, the tuples are reasonably accurate on non-native input.

3.2 Lexical and Syntactic Features

C-rater produces tuples for each document, often several per sentence. For the current experiment, we used the main verb, its subject and object. We then converted each tuple into a set of features, which included the following:

- The verb, subject (pro)noun, and object (pro)noun as individual words;
- All of the words together as a single feature;
- The verb, subject, and object words with their argument roles.

Each document can now be represented as a set of tuple-derived features, or feature vectors.

3.3 Document Comparison

Two feature vectors derived from tuples can be compared using a cosine measure (Salton 1989). The closer to 1 the cosine, the more similar the two feature sets. To compensate for different frequencies of the features and for varying document lengths, the feature vectors are weighted using standard *tf*idf* techniques.

In order to estimate the similarity between two documents, we use the following procedure. For each tuple vector in Document A, we find the tuple in Document B with the maximum cosine to the tuple in Document A. The maximum cosine values for each tuple are then averaged, resulting in a single scalar value for Document A. We call this measure *Average Maximum Cosine* (AMC).

We calculated AMCs for each student response versus the reading, the lecture, and the reading + lecture combined. This procedure was performed for both native and non-native essays. A detailed examination of the resulting trends is in section 4.

3.4 Other Measures of Document Similarity

We also performed several measures of document similarity that did not include syntactic features.

Content Vector Analysis

The student essays and the prompts were compared using Content Vector Analysis (CVA), where each document was represented as a vector consisting of the words in it (Salton 1989). The *tf*idf*-weighted vectors were compared by a cosine measure.

For non-native speakers, there was a noticeable trend. At higher score levels (where the score is determined by a human rater), student essays showed more similarity to both the reading and the lecture prompts. Both the reading and lecture similarity trends were significant (linear trend; $F = MS_{\text{linear trend}}/MS_{\text{within-subjects}} = 63$ for the reading; F=71 for the lecture at 0.05 significance level[3]). Thus, the rate of vocabulary retention from both prompts increases with higher English-language skill level.

Native speakers showed a similar pattern of increasing cosine similarity between the essay and the reading (F=35 at 0.05 significance for the trend), and the lecture (F=35 at the 0.05 level).

BLEU score

In order to measure the extent to which whole chunks of text from the prompt are reproduced in the student essays, we used the BLEU score, known from studies of machine translation (Papineni et al. 2002). We used whole essays as sections of text rather than individual sentences.

For non-native speakers, the trend was similar to that found with CVA: at higher score levels, the overlap between the essays and both prompts increased (F=52.4 at the 0.05 level for the reading; F=53.6 for the lecture).

Native speakers again showed a similar pattern, with a significant trend towards more similarity to the reading (F=35.6) and the lecture (F=31.3). These results were confirmed by a simple n-gram overlap measure.

4 Results

4.1 Overall similarity to reading and lecture

The AMC similarity measure, which relies on syntactic as well as lexical similarity, produced somewhat different results from simpler bag-of-word or n-gram measures. In particular, there was a difference in behavior between native and non-native speakers: non-native speakers showed increased structural similarity to the lecture with increasing scores, but native speakers did not.

For non-native speakers, the trend of increased AMC between the essay and the lecture was significant (F=10.9). On the other hand, there was no significant increase in AMC between non-native essays and the reading (F=3.4). Overall, for non-native speakers the mean AMC was higher for the reading than for the lecture (0.114 vs. 0.08).

Native speakers, by contrast, showed no significant trends for either the reading or the lecture. Overall, the average AMCs for the reading and the lecture were comparable (0.08 vs. 0.075).

We know from results of CVA and BLEU analyses that for both groups of speakers, higher-scoring essays are more lexically similar to the prompts. Thus, the lack of a trend for native speakers must be due to lack of increase in structural similarity between higher-scoring essays and the prompts. Since better essays are presumably better at expressing the content of the prompts, we can hypothesize that native speakers paraphrase the content more than non-native speakers.

4.2 Difference between lecture and reading

The most informative measure of speaker behavior was the difference between the Average Maximum Cosine with the reading and the lecture, calculated by subtracting the lecture AMC from the reading AMC. Here, non-native speakers showed a significant downward linear trend with increasing

[3] All statistical calculations were performed as ANOVA-style trend analyses using SPSS.

score (F=6.5; partial eta-squared 0.08), whereas the native speakers did not show any trend (F=1.5). The AMC differences are plotted in Figure 3.

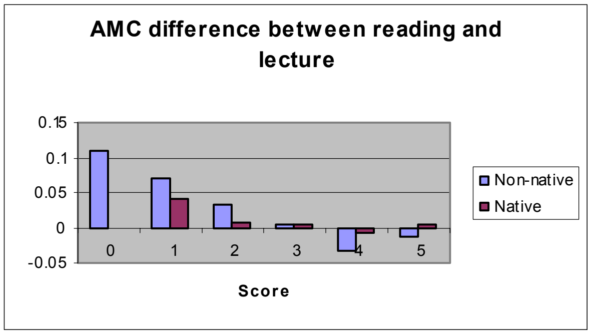

Figure 2 - AMC difference between reading and lecture

Non-native speakers with lower scores rely mostly on the reading to produce their response, whereas speakers with higher scores rely somewhat more on the lecture than on the reading. By contrast, native speakers show no correlation between score and reading vs. lecture similarity. Thus, there is a significant difference in the overall distribution and behavior between native and non-native speaker populations. This difference also shows that human raters rely on information other than simple verbatim similarity to the lecture in assigning the overall scores.

4.3 Other parameters of variation

For non-native speakers, the best predictor of the human-rated score is the difference in AMC between the reading and the lecture.

As demonstrated in the previous section, the AMC difference does not predict the score for native speakers. We analyzed native speaker demographic data in order to find any other possible predictors. The students' overall listening score, their status as monolingual vs. bilingual, their parents' educational levels all failed to predict the essay scores.

5 Discussion and Future Work

The Average Maximum Cosine measure as described in this paper successfully characterizes the behavior of native vs. non-native speaker populations on an integrated writing task. Less skillful non-native speakers show a significant trend of relying on the easier, more available prompt (the reading) than on the harder prompt (the lecture), whereas more skillful readers view the lecture as more relevant and rely on it more than on the reading. This difference can be due to better listening comprehension for the lecture and/or better memory. By contrast, native speakers rely on both the reading and the lecture about the same, and show no significant trend across skill levels. Native speakers seem to deviate more from the structure of the original prompts while keeping the same content, signaling better paraphrasing skills.

While not a direct measure of gist similarity, this technique represents a first step toward detecting paraphrases in written text. In the immediate future, we plan to extend the set of features to include non-verbatim similarity, such as synonyms and words derived by LSA-type comparison (Landauer et al. 1998). In addition, the syntactic features will be expanded to include frequent grammatical alternations such as active / passive.

A rather simple measure such as AMC has already revealed differences in population distributions for native vs. non-native speakers. Extensions of this method can potentially be used to determine if a given essay was written by a native or a non-native speaker. For instance, a statistical classifier can be trained to distinguish feature sets characteristic for different populations. Such a classifier can be useful in a number of NLP-related fields, including information extraction, search, and, of course, educational measurement.

References

Campbell, C. 1987. Writing with Others' Words: Native and Non-Native University Students' Use of Information from a Background Reading Text in Academic Compositions. Technical Report, UCLA Center for Language Education and Research.

Landauer, T.; Foltz, P. W; and Laham. D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes* 25: 259-284.

Leacock, C., & Chodorow, M. 2003. C-rater: Scoring of short-answer questions. *Computers and the Humanities,* 37(4), 389-405.

Papineni, K; Roukos, S.; Ward, T. and Zhu, W-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL* '02, p. 311-318.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, MA: Addison-Weley.