Measuring the Use of Factual Information in Test-Taker Essays

Beata Beigman Klebanov
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
bbeigmanklebanov@ets.org

Derrick Higgins
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
dhiggins@ets.org

Abstract

We describe a study aimed at measuring the use of factual information in test-taker essays and assessing its effectiveness for predicting essay scores. We found medium correlations with the proposed measures, that remained significant after the effect of essay length was factored out. The correlations did not differ substantionally between a simple, relatively robust measure vs a more sophisticated measure with better construct validity. Implications for development of automated essay scoring systems are discussed.

1 Introduction

Automated scoring of essays deals with various aspects of writing, such as grammar, usage, mechanics, as well as organization and content (Attali and Burstein, 2006). For assessment of content, the focus is traditionally on topical appropriateness of the vocabulary (Attali and Burstein, 2006; Landauer et al., 2003; Louis and Higgins, 2010; Chen et al., 2010; De and Kopparapu, 2011; Higgins et al., 2006; Ishioka and Kameda, 2006; Kakkonen et al., 2005; Kakkonen and Sutinen, 2004; Lemaire and Dessus, 2001; Rosé et al., 2003; Larkey, 1998), although recently other aspects, such as detection of sentiment or figurative language, have started to attract attention (Beigman Klebanov et al., 2012; Chang et al., 2006).

The nature of factual information used in an essay has not so far been addressed, to our knowledge; yet a misleading premise, insufficient factual basis, or an example that flies in the face of the reader's knowledge clearly detract from an essay's quality.

This paper presents a study on assessing the use of factual knowledge in argumentative essays on general topics written for a graduate school entrance exam. We propose a definition of fact, and an operationalization thereof. We find that the proposed measure has positive medium-strength correlation with essay grade, which remains significant after the impact of essay length is factored out. In order to quantify which aspects of the measure drive the observed correlations, we gradually relax the measurement procedure, down to a simple and robust proxy measure. Surprisingly, we find that the correlations do not change throughout the relaxation process. We discuss the findings in the context of validity vs reliability of measurement, and point out implications for automated essay scoring.

2 What is a Fact?

To help articulate the notion of fact, we use the following definition from a seminal text in argumentation theory: "... in the context of argumentation, the notion of fact is uniquely characterized by the idea that is held of agreements of a certain type relating to certain data, those which refer to an objective reality, and, in Poincare's words, designate essentially "what is common to several thinking beings, and could be common to all" (Perelman and Olbrechts-Tyteca, 1969, 67). Factuality is thus a matter of selecting certain kinds of data and securing a certain type of agreement over those data.

Of the different statements that refer to objective reality, the term *facts* is used to "designate ob-

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 63–72,
Montréal, Canada, June 3-8, 2012.

jects of precise, limited agreement" (Perelman and Olbrechts-Tyteca, 1969, 69). These are contrasted with *presumptions* – statements connected to what is normal and likely (*ibid.*). We suggest that the distinctions in the scope of the required agreement can be related to the referential device used in a statement: If the reference is more rigid (Kripke, 1980), that is, less prone to change in time and to indeterminacy of the boundaries, the scope of the necessary agreement is likely to be more precise and limited. With proper names prototypically being the most rigid designators, we will focus our efforts on statements about named entities.[1]

Perhaps the simplest model of the universal audience is an encyclopedia – a body of knowledge that is verified by experts, and is, therefore, "common to several thinking beings, and could be common to all" by virtue of the authority of the experts and the wide availability of the resource. However, many facts known to various groups of people that could be known to all are absent from any encyclopedia. The knowledge contained in the WWW at large, reaching not only statements explicitly contributed to an encyclopedia but also those made by people on their blogs – is perhaps as close as it gets to a working model of the universal audience.

Recent developments in Open Information Extraction make it possible to tap into this vast knowledge resource. Indeed, fact-checking is one of the applications the developers of OpenIE have in mind for their emergent technology (Etzioni et al., 2008).

3 Open Information Extraction

Traditionally, the goal of an information extraction system is automated population of structured databases of events or concepts of interest and their properties by analyzing large corpora of text (Chinchor et al., 1993; Onyshkevych, 1993; Grishman and Sundheim, 1995; Ravichandran and Hovy, 2002; Agichtein and Gravano, 2000; Davidov and Rappoport, 2009).

In contrast, the recently proposed Open Information Extraction paradigm aims to detect related pairs of entities without knowing in advance what kinds of relations exist between entities in the source data and without any seeding (Banko and Etzioni, 2008). The possibility of such extraction in English is attributed by the authors to a small number of syntactic patterns that realize binary relations between entities. In particular, they found that almost 40% of such relations are realized by the argument-verb-argument pattern (henceforth, **AVA**) (see Table 1 in Banko and Etzioni (2008)).

The TextRunner system (Banko and Etzioni, 2008) is trained using a CRF classifier on S-V-O tuples from a parsed corpus as positive examples, and tuples that violate phrasal structure as negative ones. The examples are described using features that do not require parsing or semantic role labeling. Features include part-of-speech tags, regular expressions (detecting capitalization, punctuation, etc.), context words belonging to closed classes, and conjunctions of features occurring in adjacent positions within six words of the current word.

TextRunner achieves P=0.94, R=0.65, and F-Score=0.77 on the AVA pattern (Banko and Etzioni, 2008). We note that all relations in the test sentences involve a predicate connecting two named entities, or a named entity and a date.[2] The authors kindly made available to us for research purposes a database of about 2 bln AVA extractions produced by TextRunner; this database was used in the experiments reported below.

4 Data

We randomly sampled essays written on 10 different prompts, 200 essays per prompt. Essays are graded on the scale of 1-6; the distribution of grades is shown in table 1.

Grade	**1**	**2**	**3**	**4**	**5**	**6**
%	0.6	4.9	23.5	42.6	23.8	4.7

Table 1: The distribution of grades for 2,000 essays.

[1] For example, *Barack Obama* picks out precisely one person, and the same one in 2010 as it did in 1990. In contrast, *the current US president* picks out different people every 4-8 years. For indeteminacy of boundaries, consider a statement like *US officials are wealthy*. To determine its truth, one must first secure agreement on acceptable referents of *US officials*.

[2] http://www.cs.washington.edu/research/knowitall/hlt-naacl08-data.txt

5 Building Queries from Essays

We define a query as a 3-tuple <NE,?,NP>,[3] where NE is a named entity and NP is a noun phrase from the same or neighboring sentence in a test-taker essay (the selection process is described in section 5.2). We use the pattern of predicate matches against the TextRunner database to assess the degree and the equivocality of the connection between NE and NP.

5.1 Named Entities in Test-Taker Essay

We use the Stanford Named Entity Recognizer (Finkel et al., 2005) that tags named entities as people, locations, organizations, and miscellaneous. We annotated a sample of 90 essays for named entities; the sample yielded 442 tokens, which we classified as shown in Table 2. The Enamex classes (people, locations, organizations) account for 58% of all the entities in the sample. The recognizer's recall of people and locations is excellent (though they are not always classified correctly – see caption of Table 2), although test-taker essays feature additional entity types that are not detected as well.

Category	Recall	Examples
Location	0.98	Iraq, USA
Person	0.96	George W. Bush, Freud
Org.	0.87	Guggenheim Foundation
Gov.	0.79	No Child Left Behind
Awards	0.79	Nobel Prize
Events	0.68	Civil War, World War I
Sci & Tech	0.59	GPS, Windows 3.11
Art	0.44	Beowulf, Little Women

Table 2: Recall of the Stanford NER by category. Note that an entity is counted as recalled as long as it is identified as belonging to any NE category, even if it is misclassified. For example, *Freud* is tagged as location, but we count it towards the recall of people.

In terms of precision, we observed that the tagger made few clear mistakes, such as tagging sentence-initial adverbs and their mis-spelled versions as named entities (*Eventhough*, *Afterall*). The bulk of the 96 items over-generated by the tagger are in the "grey area" – while we haven't marked them, they are not clearly mistakes. A common case are names of national and religious groups, such as *Muslim* or *Turkish*, or capitalizations of otherwise common nouns for emphasis and elevation, such as *Arts* or *Masters*. Given our objective to ground the queries in items with specific referents, these are less suitable. If all such cases are counted as mistakes, the tagger's precision is 82%.

5.2 Selection of NPs

We employ a grammar-based approach for selecting NPs. We use the Stanford dependency parser (de Marneffe et al., 2006; Klein and Manning, 2003) to determine dependency relations.

In order to find out which dependency paths connect between named entities and clearly related NPs in essays, we manually marked concepts related to 95 NEs in 10 randomly sampled essays. We marked 210 query-able concepts in total. The resulting 210 dependency paths were classified according to the direction of the movement.

Out of the 210 paths, 51 (24%) contain a single upward or downard step, that is, are cases where the NE is the head of the constituent in which the NP is embedded, or the other way around. Some examples are shown in Figure 1. Note that the predicate connecting NE and NP is not lexicalized, but the existence of connection is signaled by the close-knit grammatical pattern.

The most prolific family of paths starts with an upward step, followed by a sequences of 1-4 downwards steps; 71 (34%) of all paths are of this type. Most typically, the first upward move connects the NE to the predicate of which it is an argument, and, down from there, to either the head of another argument (↑↓) or to an argument's head's modifier (↑↓↓). These are explicit relations, where the relation is typically lexicalized by the predicate.

We expand the context of extraction beyond a single sentence only for NEs classified as PERSON. We apply a gazetteer of private names by gender from US Census 2010 to expand a NE of a given gender with the appropriate personal pronouns; a word that is a part of the original name (only surname, for

[3] We do not attempt matching the predicate, as (1) in many cases there is no clearly lexicalized predicate (see the discussion of single step patterns in section 5.2) and (2) adding a predicate field would make matches against the database sparser (see section 6.1).

[4] NE=Kroemer; NP=Heterojunction Bipolar Transitor

↓ a Nobel Prize in a science field

↓ Chaucer, in the 14 century, ...

↑ the prestige of the Nobel Prize

↑ Kidman's talent

↑↓ Kroemer received the Nobel Prize

↑↓↓ Kroemer received the Nobel Prize for his work on the Heterojunction Bipolar Transitor[4]

Figure 1: Examples of dependency paths used for query construction.

example), is also considered an anaphor and a candidate for expansion. We expand the context of the PERSON entity as long as the subsequent sentence uses any of the anaphors for the name. This way, we hope to capture an extended discussion of a named entity and construct queries around its anaphoric mentions just as we do around the regular, NE mention. A name that is not predominantly male or female is not expanded with personal pronouns. Table 3 shows the distribution of queries automatically generated from the sample of 2,000 essays.

↑	2,817	15.9%
↓	798	4.5%
↑↑	813	4.6%
↓↓	372	2.1%
↑↓	4,940	27.8%
↑↓↓	2,691	15.1%
↑↓↓↓	1,568	8.8%
↑↑↓	3,772	21.2%
total	17,771	100%

Table 3: Distribution of queries by path type.

6 Matching and Filtering Queries

6.1 Relaxation for improved matching

To estimate the coverage of the fact repository with respect to the queries extracted from essays, we submit each query to the TextRunner repository in the <NE,?,NP> format and record the number of times the repository returned any matches at all. The percentage of matched queries is 21%. To increase the chances of finding a match, we process the NP to remove determiners and pre-modifiers of the head that are very frequent words, such as removing *a very* from *a very beautiful photograph*.

Additionally, we produce three variants of the NP. The first, NP_1, contains only the sequence of nouns ending with the head noun; in the example, NP_1 would be *photograph*. The second variant, NP_2, contains only the word that is rarest in the whole of NP. All capitalized words are given the lowest frequency of 1. Thus, if any of the NP words are capitalized, the NP_2 would either contain an out of vocabulary word to the left of the first capitalized word, or the leftmost capitalized word. This means that names would typically be split such that only the first name is taken. For example, the NP *the author Orhan Phamuk* would generate NP_2 *Orhan*. When no capitalized words exist, we take the rarest one, thus a NP *category 3 hurricane* would yield NP_2 *hurricane*. The third variant only applies to NPs with capitalized parts, and takes the rightmost capitalized word in the query. Thus, the NP *the actress Nicole Kidman* would yield NP_3 *Kidman*.

Applying these procedures to every NP inflates the number of actual queries posed to the TextRunner repository by almost two-fold (31,211 instead of 17,771), while yielding a 50% increase in the number of cases where at least one variant of the original query had at least one match against the repository (from 21% to 35%).

6.2 Match-specific filters

In order to zero in on matches that correpond to factual statements and indeed pertain to the queried arguments, we implement a number of filters.

Predicate filters

We filter out modal and hedged predicates, using lists of relevant markers. We remove predicates like *might turn out to be* or *possibly attended*, as well as future tense predicates (marked with *will*).

Argument filters

For matches that passed the predicate filters, we check the arguments. Let **mARG** be the actual string that matched ARG (ARG $\in$\{NE,NP\}). Let **EC** (Essay Context) refer to source sentence(s) in

the essay.[5] We filter out the following matches:

- Capitalized words follow ARG in mARG that are not in EC;
- >1 capitalized or rare words precede ARG in mARG that are not in EC and not honorifics;
- mARG is longer than 8 words;
- More than 3 words follow ARG in mARG.

The filters target cases where mARG is more specific than ARG, and so the connection to ARG might be tenuous, such as ARG=*Harriet Beecher Stowe*, mARG = *Harriet Beecher Stowe Center*.

6.3 Filters based on overall pattern of matches

6.3.1 Negation filter

For all matches for a given query that passed the filters in section 6.2, we tally positive vs negative predicates.[6] If the ratio of negative to positive is above a threshold (we use 0.1), we consider the query an unsuitable candidate for being "potentially common to all," and therefore do not credit the author with having mentioned a fact.

This criterion of potential acceptance by a universal audience fails a query such as <Barack Obama,?,US citizen>, based on the following pattern of matches:

Count	Predicate
10	is not
4	is
2	was always
1	is really
1	isn't
1	was not

In a similar fashion, an essay writer's statement that "The beating of Rodney King in Los Angeles ... made for tense race relations" is not quite in accord with the 16 hits garnered by the statement "The Los Angeles riots were not caused by the Rodney King verdict," against other hits with predicates like *erupted after*, *occurred after*, *resulted from*, *were sparked by*, *followed*.

[5] A single sentence, unless anaphor-based expansion was carried out; see section 5.2.

[6] We use a list of negation markers to detect those.

Somewhat more subtly, the connection between *Albert Einstein* and *atomic bomb*, articulated as "For example, Albert Einstein's accidental development of the atomic bomb has created a belligerent technological front" by a test-taker, is opposed by 6 hits with the predicate *did not build* against matches with predicates such as *paved the way to*, *led indirectly to*, *helped in*, *created the theory of*. The conflicting accounts seem to reflect a lack of consensus on the degree of Einstein's responsibility.

The cases above clearly demonstrate the implications of the *argumentative* notion of facts used in our project. Facts are statements that the audience is prepared to accept without further justification, differently from arguments, and even from presumptions (statements about what is normal and likely), for which, as Perelman and Olbrechts-Tyteca (1969) observe, "additional justification is beneficial for strengthening the audience's adherence." Certainly in the Obama case and possibly in others, a different notion of factuality, for example, a notion that emphasizes availability of legally acceptable supporting evidence, would have led to a different result. Yet, in an ongoing instance of argumentation, the mere *need* to resort to such a proof is already a sign that the audience is not prepared to accept a statement as a fact.

6.4 Additional filters

We also implemented a number of filters aimed at detecting excessive diversity in the matches, which could suggest that there is no clear and systematic relation between the NE and the NP. The filters are conjunctions of thresholds operating over measures such as purity of matches (percentage of exact matches in NE or NP), degree of overlap of non-pure matches with the context of the query in the essay, clustering of the predicates (recurrence of the same predicates across matches), general frequencies of NE and NP.

7 Evaluation

7.1 Manual check of queries

A manual check of a small subset of queries was initially intended as an interim evaluation of the query construction process, to see how often the produced queries are deficient candidates for later verification.

However, we also decided to include a human fact-check of the queries that were found to be verifiable, to see the kinds of factual mistakes made in essays.

A research assistant was asked to classify 500 queries into **W**rong (the NE and NP are not related in the essay), **T**rivial (almost any NE could be substituted, as in <WWI,?, Historians>), **S**ubjective (<T.S.Eliot,?,the most frightening poet of all time>), **VC** – verifiable and correct, **VI** – verifiable and incorrect. Table 4 shows the distribution.

W	T	S	VC	VI
18%	13%	13%	54%	2%

Table 4: The distribution of query types for 500 queries.

Queries classified as Wrong (18%) mostly correspond to parser mistakes. Trivial and Subjective queries, while not attributing to the author connections that she has not made, are of questionable value as far as fact-checking goes. Perhaps the most surprising figure is the meager amount of verifiable and incorrect queries. Examples of relevant statements from essays include (NE and NP are boldfaced):

- For example, **Paul Gaugin** who was a **successful business man**, with a respectable wife and family, suddenly gave in to the calling of the arts and left his life. (He was a *failing* businessman immediately before leaving family.)
- For example, in **Jane Austin**'s **Little Women**, she portrays the image of a lovely family and the wonders of womenhood. (The book is by Louisa May Alcott.)
- This occurrence can be seen with the **Rodney King** problem in California during the **late 1980's**. (The Rodney King incident occurred on March 3, 1991).
- We see the philosophers Aristotle, Plato, **Socrates** and their **practical writings** of the political problems and issues of the day. (Socrates is not known to have left writings.)

First, we observe that factual mistakes are rare. Furthermore, they seem to pertain to one in a series of related facts, most of which are correct and testify to the author's substantial knowledge about the matter – consider Paul Gaugin's biography or the contents of "Little Women" in the examples above. It is therefore unclear how detrimental the occasional factual "glitches" are to the quality of the essay.

8 Application to Essay Scoring

We show Pearson correlations between human scores given to essays and a number of characteristics derived from the work described here, as well as the partial correlations when the effect of essay length is factored out. We calculated both the correlations using raw numbers and on a logarithmic scale, with the latter generally producing higher correaltions. Therefore, we are reporting the correlations between grade and the logarithm of the relevant characteristic. The characteristics are:

#NE The number of NE tokens in an essay.

#Queries The number of queries generated by the system from the given essay (as described in section 5.2).

#Matched Queries The number of queries for which a match was found in the TextRunner database. If the original query or any of its expansion variants (see section 6.1) had matches, the query contributes a count of 1.

#Filtered Matches The number of queries that passed the filters introduced in section 6. If the original query or any of its expansion variants passed the filters, the query contributes a count of 1.

Table 5 shows the results. First, we find that all correlations are significant at p=0.05, as well as the partial correlations exluding the effect of length for 7 out of 10 prompts. All correlations are positive, that is, the more factual information a writer employs in an essay, the higher the grade – beyond the oft reported correlations between the grade and the length of an essay (Powers, 2005).

Second, we notice that all characteristics – from the number of named entities to the number of filtered matches – produce similar correlation figures.

Third, there are large differences between average numbers of named entities per essay across prompts.

Prompt	NE	Pearson Corr. with Grade				Partial Corr. Removing Length			
		#NE	#Q	#Mat.	# Filt.	#NE	#Q	#Mat.	# Filt.
P1	280	0.144	0.154	0.182	0.185	0.006	0.019	0.058	0.076
P2	406	0.265	0.259	0.274	0.225	0.039	0.053	0.072	0.069
P3	452	0.245	0.225	0.188	0.203	0.049	0.033	0.009	0.051
P4	658	0.327	0.302	0.335	0.327	0.165	0.159	0.177	0.160
P5	704	0.470	0.477	0.473	0.471	0.287	0.294	0.304	0.305
P6	750	0.429	0.415	0.388	0.373	0.271	0.242	0.244	0.257
P7	785	0.470	0.463	0.479	0.469	0.302	0.302	0.341	0.326
P8	838	0.423	0.390	0.406	0.363	0.264	0.228	0.266	0.225
P9	919	0.398	0.445	0.426	0.393	0.158	0.209	0.233	0.219
P10	986	0.455	0.438	0.375	0.336	0.261	0.257	0.170	0.175
AV.	678	0.363	0.357	0.353	0.335	0.180	0.180	0.187	0.186

Table 5: Pearson correlation and partial correlation removing the effect of length between a number of characteristics (all on a log scale) and the grade. The second column shows the total number of identified named entities in the 200-essay sample from the given prompt. The prompts are sorted by the second column.

Generally, the higher the number, the better the number of named entities in the essay predicts its grade (the more NEs the higher the grade). This suggests that the use of named entities might be relatively irrelevant for some prompts, and much more relevant for others. For example, prompt P10 reads "The arts (painting, music, literature, etc.) reveal the otherwise hidden ideas and impulses of a society," thus practically inviting exemplification using specific works of art or art movements, while success with prompt P1 – "The human mind will always be superior to machines because machines are only tools of human minds" – is apparently not as dependent on named entity based exemplification. Excluding prompts with smaller than average total number of named entities (<678), the correlations average 0.40-0.44 across the various characteristics, with partial correlations averaging 0.25-0.26.

9 Discussion and Conclusion

9.1 Summary of the main result

In this article, we proposed a way to measure the use of factual information in text-taker essays. We demonstrated that the use of factual information is indicative of essay quality, observing positive correlations between the count of instances of fact-use in essays and the grade of the essay, beyond what can be attributed to a correlation between the total number of words in an essay and the grade.

9.2 What is driving the correlations?

We also investigated which of the components of the fact-use measure were responsible for the observed correlations. Specifically, we considered (a) the number instances of fact-use that were verified against a database of human-produced assertions, filtered for controversy and excessive diversity; (b) the number of instances of fact-use that were verified against the database, without subsequent filtering; (c) the number of instances of fact-use identified in an essay (without checking against the database); (d) the number of named entities used in an essay (without constructing queries around the entity). These steps correspond to a gradual relaxation of the full fact-checking procedure all the way to a proxy measure that counts the number of named entities.

We observed similar correlations throughout the relaxation procedure. We therefore conclude that the number of named entities is the driving force behind the correlations, with no observed effect of the query construction and verification procedures.[7] This result could be explained by two factors.

First, a manual check of 500 queries showed that factual mistakes are rare – only 2% of the queries corresponded to factually incorrect statements. Furthermore, mistakes were often accompanied by the

[7] While the trend is in the direction of an increase in Pearson correlations from (a) to (d), the differences are not statistically significant.

test-taker's use of additional facts about the same entity which were correct; this might alleviate the impact of a mistake in the eyes of a grader.

Second, the query verification procedure applied to only about 35% of the queries – those for which at least one match was found in the database, that is, 65% of the queries could not be assessed using the database of 2 bln extractions. The verification procedure is thus much less robust than the procedure for detecting named entities, which performs at above $>$80% recall and precision.

9.3 Implications for automated scoring

Our results suggest that essays on a general topic written by adults for a high-stakes exam contain few incorrect facts, so the potential for a full fact-checking system to improve correlations with grades beyond merely detecting the potential for a factual statement using a named entity recognizer is not large. While a measure based on the number of "verified" facts found in an essay demonstrated a significant correlation with human scores beyond the contribution of essay length, a simpler measure based only on the number of named entities in the essay demonstrated a similar relationship with human scores.

Given the similarity in the two features' empirical usefulness, it would seem that the feature that counts the number of named entities in an essay is a better candidate, due to its simplicity and robustness. However, there is another perspective from which a feature based only on the number of named entities in an essay may be less suitable for use in scoring: the perspective of *construct validity*, the degree to which a test (or, in this case, a scoring system) actually measures what it purports to. As mentioned above, the number of named entities in an essay is, at best, a proxy measure,[8] roughly indicative of the referencing of factual statements in support of an argument within an essay. Because the measure itself is not directly sensitive to how named entities are used in the essay, though, even entities with no connection to the essay topic would tend to contribute to the score, and the measure is therefore vulnerable to manipulation by test-takers.

[8]For a discussion of *proxes* vs *trins* in essay grading, see (Page and Petersen, 1995).

An obvious strategy to exploit this scoring mechanism would be to simply include more named entities in an essay, either interspersing them randomly throughout the text, or including them in long lists of examples to illustrate a single point. Such a blatant approach could potentially be detected by the use of a filter or *advisory* (Higgins et al., 2006; Landauer et al., 2003) designed to identify anomalous writing strategies. However, there could be more subtle approaches to exploiting such a feature. For example, it is possible that test-takers might be inclined to increase their use of named entities by adducing more facts in support of an argument, and would go beyond the comfort zone of their actual factual knowledge, thus making more factual mistakes. Test gaming strategies have been recognized as a threat to automated scoring systems for some time (Powers et al., 2001), and there is evidence based on test takers' own self-reported behavior that this threat is real (Powers, 2011). This is one major reason why large-scale operational testing programs (such as GRE or TOEFL) use automated essay scoring only in combination with human ratings. In sum, the degree to which a linguistic feature is predictive of human essay scores is not the only criterion for evaluation; the washback effects of using the feature (on writing behavior and on instruction) must also be considered.

The second finding of this study is that the effectiveness of fact-checking for essay assessment is compromised by the limited coverage of the wealth of factual statements made by essay writers, with only 35% of queries garnering any hits at all in a large general-purpose database of assertions. It is possible, however, that OpenIE technology can be used to collect more focused repositories on specific topics, such as the history of the American Civil War, which could be used to assess responses to tasks related to that particular subject matter. This is one of the directions of our future research.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM conference on Digital Libraries*, pages 85–94. ACM.

Yigal Attali and Jill Burstein. 2006. Automated Es-

say Scoring With e-rater®V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Michele Banko and Oren Etzioni. 2008. The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 28–36, Columbus, OH, June. Association for Computational Linguistics.

Beata Beigman Klebanov, Jill Burstein, Nitin Madnani, Adam Faulkner, and Joel Tetreault. 2012. Building Subjectivity Lexicon(s) From Scratch For Essay Data. In *Proceedings of CICLING*, New Delhi, India.

Tao-Hsing Chang, Chia-Hoang Lee, and Yu-Ming Chang. 2006. Enhancing Automatic Chinese Essay Scoring System from Figures-of-Speech. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 28–34.

Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, and Tao-Hsing Chang. 2010. An Unsupervised Automated Essay Scoring System. *IEEE Transactions on Intelligent Systems*, 25(5):61–67.

Nancy Chinchor, Lynette Hirschman, and David Lewis. 1993. Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3):409–449.

Dmitry Davidov and Ari Rappoport. 2009. Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns. In *Proceedings of EMNLP*, pages 267–275.

Arijit De and Sunil Kopparapu. 2011. An unsupervised approach to automated selection of good essays. In *Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE*, pages 662 –666.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC*, pages 449–454, Genoa, Italy, May.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Jenny Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, MI, June. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Proceedings of MUC*, pages 1–11.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145–159.

Tsunenori Ishioka and Masayuki Kameda. 2006. Automated Japanese Essay Scoring System based on Articles Written by Experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 233–240, Sydney, Australia, July. Association for Computational Linguistics.

Tuomo Kakkonen and Erkki Sutinen. 2004. Automatic assessment of the content of essays based on course materials. In *Proceedings of the International Conference on Information Technology: Research and Education*, pages 126–130, London, UK.

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen. 2005. Automatic Essay Grading with Probabilistic Latent Semantic Analysis. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 29–36, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Dan Klein and Christopher Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July. Association for Computational Linguistics.

Saul Kripke. 1980. *Naming and Necessity*. Harvard University Press.

Thomas Landauer, Darrell Laham, and Peter Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Mark Shermis and Jill Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Leah Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of SIGIR*, pages 90–95, Melbourne, AU.

Benoît Lemaire and Philippe Dessus. 2001. A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, 24:305–320.

Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95, Los Angeles, California, June. Association for Computational Linguistics.

Boyan Onyshkevych. 1993. Template design for information extraction. In *Proceedings of MUC*, pages 19–23.

Ellis Page and Nancy Petersen. 1995. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, 76:561–565.

Chaïm Perelman and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric: A Treatise on Argumentation*. Notre

Dame, Indiana: University of Notre Dame Press. Translated by John Wilkinson and Purcell Weaver from French original published in 1958.

Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich. 2001. Stumping E-Rater: Challenging the Validity of Automated Essay Scoring. *ETS research report RR-01-03*, http://www.ets.org/research/policy_research_reports/rr-01-03.

Donald Powers. 2005. "Wordiness": A selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses. *ETS research memorandum RM-04-08*, http://www.ets.org/research/policy_research_reports/rm-04-08.

Donald Powers. 2011. Scoring the TOEFL Independent Essay Automatically: Reactions of Test Takers and Test Score Users. *ETS research manuscript RM-11-34*, http://www.ets.org/research/policy_research_reports/rm-11-34.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a Question Answering System. In *Proceedings of ACL*, pages 41–47.

Carolyn Rosé, Antonio Roqueand, Dumisizwe Bhembe, and Kurt VanLehn. 2003. A hybrid text classification approach for analysis of student essays. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 29–36.