Word Association Profiles and their Use for Automated Scoring of Essays

Beata Beigman Klebanov and Michael Flor
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541
{bbeigmanklebanov,mflor}@ets.org

Abstract

We describe a new representation of the content vocabulary of a text we call *word association profile* that captures the proportions of highly associated, mildly associated, unassociated, and dis-associated pairs of words that co-exist in the given text. We illustrate the shape of the distirbution and observe variation with genre and target audience. We present a study of the relationship between quality of writing and word association profiles. For a set of essays written by college graduates on a number of general topics, we show that the higher scoring essays tend to have higher percentages of both highly associated and dis-associated pairs, and lower percentages of mildly associated pairs of words. Finally, we use word association profiles to improve a system for automated scoring of essays.

1 Introduction

The vast majority of contemporary research that investigates statistical properties of language deals with characterizing *words* by extracting information about their behavior from large corpora. Thus, co-occurrence of words in n-word windows, syntactic structures, sentences, paragraphs, and even whole documents is captured in vector-space models built from text corpora (Turney and Pantel, 2010; Basili and Pennacchiotti, 2010; Erk and Padó, 2008; Mitchell and Lapata, 2008; Bullinaria and Levy, 2007; Jones and Mewhort, 2007; Pado and Lapata, 2007; Lin, 1998; Landauer and Dumais, 1997; Lund and Burgess, 1996; Salton et al., 1975). However, little is known about typical profiles of *texts* in terms of co-occurrence behavior of their words. Some information can be inferred from the success of statistical techniques in predicting certain structures in text. For example, the fact that a text segmentation algorithm that uses information about patterns of word co-occurrences can detect sub-topic shifts in a text (Riedl and Biemann, 2012; Misra et al., 2009; Eisenstein and Barzilay, 2008) tells us that texts contain some proportion of more highly associated word pairs (those in subsequent sentences within the same topical unit) and of less highly associated pairs (those in sentences from different topical units).[1] Yet, does each text have a different distribution of highly associated, mildly associated, unassociated, and dis-associated pairs of words, or do texts tend to strike a similar balance of these? What are the proportions of the different levels of association, how much variation there exists, and are there systematic differences between various kinds of texts? We present research that makes a first step in addressing these questions.

From the applied perspective, our interest is in quantifying differences between well-written and poorly written essays, for the purposes of automated scoring of essays. We therefore concentrate on essay data for the main experiments reported in this paper, although some additional corpora will be used for illustration purposes.

The paper is organized as follows. Section 2 presents our methodology for building word association profiles for texts. Section 3 illustrates the profiles for three corpora from different genres. Section 4.2 presents our study of the relationship between writing quality and patterns of word associations, with section 4.5 showing the results of adding a feature based on word association profile to a state-of-art essay scoring system. Related work is reviewed is section 5.

[1] Note that the classical approach to topical segmentation of texts, TextTiling (Hearst, 1997), uses only word repetitions. The cited approaches use topic models that are in turn estimated using word co-occurrence.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1148–1158,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

2 Methodology

In order to describe the word association profile of a text, three decisions need to be made. The first decision is how to quantify the extent of co-occurrence between two words; we will use point-wise mutual information (**PMI**) estimated from a large and diverse corpus of texts. The second is which pairs of words in a text to consider when building a profile for the text; we opted for all pairs of content word types occurring in a text, irrespective of the distance between them. We consider word types, not tokens; no lemmatization is performed. The third decision is how to represent the co-occurrence profiles; we use a histogram where each bin represents the proportion of word pairs in the given interval of PMI values. The rest of the section gives more detail about these decisions.

To obtain comprehensive information about typical co-occurrence behavior of words of English, we build a first-order co-occurrence word-space model (Turney and Pantel, 2010; Baroni and Lenci, 2010). The model was generated from a corpus of texts of about 2.5 billion words, counting co-occurrence in a paragraph,[2] using no distance coefficients (Bullinaria and Levy, 2007). About 2 billion words come from the Gigaword 2003 corpus (Graff and Cieri, 2003). Additional 500 million words come from an in-house corpus containing popular science and fiction texts. Occurrence counts of 2.1 million word types and of 1,279 million word type pairs are efficiently compressed using the TrendStream technology (Flor, 2013), resulting in a database file of 4.7GB. TrendStream is a trie-based architecture for storage, retrieval, and updating of very large word n-gram datasets. We store pairwise word associations as bigrams; since associations are unordered, only one of the orders in actually stored in the database.

There is an extensive literature on the use of word-association measures for NLP, especially for detection of collocations (Pecina, 2010; Evert, 2008; Futagi et al., 2008). The use of point-wise mutual information with word-space models is noted in (Zhang et al., 2012; Baroni and Lenci, 2010; Mitchell and Lapata, 2008; Turney, 2001). Point-wise mutual information is defined as follows (Church and Hanks, 1990):

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

Differently from Church and Hanks (1990), we disregard word order when computing $P(x, y)$. All probabilities are estimated using frequencies.

We define $\mathbf{WAP_T}$ – a **word association profile** of a text T – as the distribution of PMI(x, y) for all pairs of content[3] word types (x, y) $\in$T. All pairs of word types for which the associations database returned a null value (the pair has never been observed in the same paragraph) are excluded from the calculation. For our main dataset (described later as setA, section 4.1), the average percentage of non-null values per text is 92%.

To represent the WAP of a text, we use a 60-bin histogram spanning all PMI values. The lowest bin (shown in Figures 1 and 2 as PMI = –5) contains pairs with PMI$\leq$–5; the topmost bin (shown in Figures 1 and 2 as PMI = 4.83) contains pairs with PMI $>$ 4.67, while the rest of the bins contain word pairs (x, y) with -5 <PMI(x, y) $\leq$ 4.67. Each bin in the histogram (apart from the top and the bottom ones) corresponds to a PMI interval of 0.167. We chose a relatively fine-grained binning and performed no optimization for grid selection; for more sophisticated gridding approaches to study non-linear relationships in the data, see Reshef et al. (2011).

We will say that a text A is **tighter** than text B if the WAP of A is shifted towards the higher end of PMI values relative to text B. The intuition behind the terminology is that texts with higher proportions of highly associated pairs are likelier to be more focused, dealing with a small number of topics at greater length, as opposed to texts that bring various different themes into the text to various extents. Thus, the text "The dog barked and wagged its tail" is much tighter than the text "Green ideas sleep furiously", with all the six content word pairs scoring above PMI=5.5 in the first and below PMI=2.2 in the second.[4]

3 Illustration: The shape of the distribution

For a first illustration, we use a corpus of 5,904 essays written as part of a standardized graduate

[2] In all texts, we use human-marked paragraphs, indicated either by a new line or by an xml markup.

[3] We part-of-speech tag a text using OpenNLP tagger (http://opennlp.apache.org) and only take into account common and proper nouns, verbs, adjectives, and adverbs.

[4] We omitted *colorless* from the second example, as *colorless* is actually highly associated with *green* (PMI=4.36).

school admission test (a full descrption of these data is given in section 4.1, under setA p1-p6). For each essay, we compute the WAP and represent it using the 60-bin histogram. For each bin in the histogram, we compute its average value over the 5,904 essays; additionally, we compute the 15^{th} and 85^{th} percentiles for each bin, so that the band between them contains values observed for 70% of the texts. The series with the solid thick (blue) line in Figure 1 shows the distribution of the average percentage of word type pairs per bin (essays-av); the dotted lines above and below show the band capturing the middle 70% of the distribution (essays-15 and essays-85).

We observe that the shape of the WAP is very stable across essays, and the variation around the average is quite limited.

Next, consider the thin solid (green) line with asterisk-shaped markers in Figure 1 that plots a similarly-binned histogram for the normal distribution with μ=0.90 and σ=0.66. We note that for values below PMI=2.17, the normal curve is within or almost within the 70% band for the essay data. The divergence occurs at the right tail with PMI>2.17, that covers, on average, about 8% of the pairs (5.6% and 10.4% for the 15^{th} and 85^{th} percentiles, respectively).

To get an idea about possible variation in the distribution, we consider two additional corpora from different genres. We use a corpus of Wall Street Journal 1987 articles from the TIPSTER collection.[5] We picked articles of 250 to 700 words in length, in order to keep the length of texts comparable to the essay data, while varying the genre; 770 such articles were found. The dashed (orange) line in Figure 1 shows the distribution of average values for the WSJ collection (wsj-av). We observe that the shape of the distribution is similar to that of essay data, although WSJ articles tend to be less tight, on average, since the distribution in PMI<2.17 area in the WSJ data is shifted to the left relative to essays. Yet, the picture at the right tail is remarkably similar to that of the essays, with 9% of word pairs, on average, having PMI>2.17.

The second additional corpus contains 140 literary texts written or adapted for readers in grades 3 and 4 in US schools (Sheehan et al., 2008). In terms of length, these texts fall into the same range as the other corpora, averaging 507 words. The average WAP for these texts is shown with a thin solid (purple) line with circular markers in Figure 1 (Grades 3-4). These texts are much tighter than texts in the other two collections, as the distribution is shifted to the right. The right tail, with PMI>2.17, holds 19% of all word pairs in these texts – more than twice the proportion in essays written by college graduates or in texts from the WSJ.

It is instructive to check whether the over-use of highly associated pairs is *felt* during reading. These texts strike an adult reader as overly explicit, taking the space to state things that an adult reader would readily infer or assume. For example, consider the following opening paragraph:

> "Grandma Rose gave Daniel a recorder. A recorder is a musical instrument. Daniel learned to play by blowing on the recorder. It didn't take lots of air. It didn't take big hands to hold since it was pocket-sized. His fingers covered the toneholes just fine. Soon Daniel played entire songs. His mother loved to listen. Sometimes she hummed along with Daniel's recorder."

The second and the third sentences state things that for an adult reader would be too obvious to need mention. In fact, these sentences almost seem like *training* sentences – the kind of sentences from which the associations between *recorder* and *musical instrument*, *play*, *blowing* can be learned. According to Hoey's theory of lexical priming (Hoey, 2005), one of the main functions of schooling is to imbue children with the societally sanctioned word associations.

To conclude the illustration, we observe that there are some broad similarities between the different copora in terms of the distribution of pairs of word types. Thus, texts seem to be mainly made of pairs of weakly associated words – about half the pairs of word types lie between PMI of 0.5 and 1.5, in all the examined collections (52% for essays, 44% for each of WSJ and young reader corpora). The percentages of pairs at the low and the high ends of PMI differ with genre – writing for children favors the higher end, while typical Wall Street Journal writing favors the low end, relatively to a corpus of essays on general topics written by college graduates.

These observations are necessarily very tentative, as only a few corpora were examined. Still,

[5] LDC93T3A in LDC catalogue

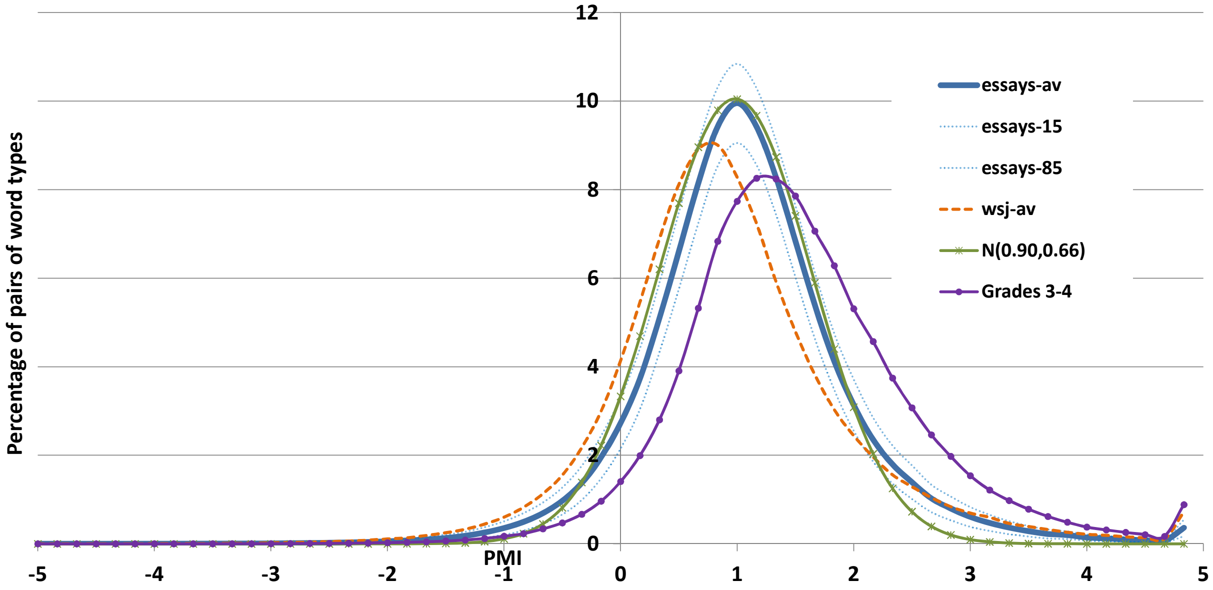

Figure 1: WAP histograms for three corpora, shown with smooth lines instead of bars for readability. Average for essays (a thick solid blue line), average for WSJ articles (a dashed orange line); average for Grades 3-4 corpus (a thin solid purple line with round markers). Normal distribution is shown with a thin solid green line with asterisk markers. Middle 70% of essays fall between the dotted lines.

we believe the illustration is suggestive, in that there is both constancy in writing for a similar purpose (observe the limited variation around the average that captures 70% of the essays) and variation with genre and target audience. In what follows, we will explore more thoroughly the information provided by word association profiles regarding the quality of writing.

4 Application to Essay Scoring

Texts written for a test and scored by relevant professionals is a setting where variation in text quality is expected. In this section, we report our experiments with using WAPs to explore the variation in quality as quantified by essay scores. We first describe the data (section 4.1), then show the patterns of relationships between essay scores and word association profiles (section 4.2). Finally, we report on an experiment where we significantly improve the performance of a very competitive, state-of-art system for automated scoring of essays, using a feature derived from WAP.

4.1 Data

We consider two collections of essays written as responses in an analytical writing section of a high-stakes standardized test for graduate school admission; the time limit for essay composition was 45 minutes. Essays were written in response to a prompt (essay question). A prompt is usually a general statement, and the test-taker is asked to develop an argument supporting or refuting the statement. Example prompts are: “High-speed electronic communications media, such as electronic mail and television, tend to prevent meaningful and thoughtful communication” and “In the age of television, reading books is not as important as it once was. People can learn as much by watching television as they can by reading books.”

The first collection (henceforth, **setA**) contains 8,899 essays written in response to nine different prompts, about 1,000 per prompt;[6] the per-prompt subsets will be termed **setA-p1** through **setA-p9**. Each essay in setA was scored by 1 to 4 human raters on a scale of 1 to 6; the majority of essays received 2 human scores. We use the average of the available human scores as the gold-standard score for the essay. Most essays thereby receive an integer score,[7] so the ranking of the essays is coarse. From this set, p1-p6 were used for feature selection, data visualization, and estimation of the regression models (training), while sets p7-p9 were reserved for a blind test.

The second collection (henceforth, **setB**) con-

[6] While we sampled exactly 1,000 essays per prompt, we removed empty responses, resulting in 975 to 1,000 essays per sample.

[7] as the two raters agree most of the time

tains 400 essays, with 200 essays written on each of two prompts given as examples above (**setB-p1** and **setB-p2**). In an experimental study by Attali et al. (2013), each essay was scored by 16 professional raters on a scale of 1 to 6, allowing plus and minus scores as well, quantified as 0.33 – thus, a score of 4- is rendered as 3.67. This fine-grained scale resulted in higher mean pairwise inter-rater correlations than the traditional integer-only scale (r=0.79 vs around r=0.70 for the operational scoring). We use the average of 16 raters as the final grade for each essay. This dataset provides a very fine-grained ranking of the essays, with almost no two essays getting exactly the same score.

Rounded	setA p1-p9			setB	
Score	av	min	max	p1	p2
1	.01	.00	.01	–	–
2	.05	.04	.06	.03	.03
3	.25	.20	.29	.30	.28
4	.44	.42	.47	.54	.55
5	.21	.16	.24	.13	.14
6	.04	.02	.07	.01	.02

Table 1: Score distribution in the essay data. For the sake of presentation in this table, all scores were rounded to integer scores, so a score of 3.33 was counted as 3, and a score of 3.5 was counted as 4. A cell with the value of .13 (row titled 5 and column titled SetB p1) means that 13% of the essays in setB-p1 received scores that round to 5. For setA, average, minimum, and maximum values across the nine prompts are shown.

Table 1 shows the distribution of rounded scores in both collections. Average essay scores are between 3.74 to 3.98 across the different prompts from both collections. The use of 16 raters seems to have moved the rounded scores towards the middle; however, the relative ranking of the essays is much more delicate in setB than in setA.

4.2 Essay Score vs WAP

We calculated correlations between essay score and the proportion of word pairs in each of the 60 bins of the WAP histogram, separately for each of the prompts p1-p6 in setA. For a sample of 1,000 instances, a correlation of r=0.065 is significant at $p = 0.05$. Figure 2 plots the correlations.

First, we observe that, perhaps contrary to expectation, the proportion of the highest values of PMI (the area to the right of PMI=4 in Figure 2) does not yield a consistent correlation with essay scores. Thus, inasmuch as highest PMI values tend to capture multi-word expressions (*South* and *Africa*; *Merill* and *Lynch*), morphological variants (*bids* and *bidding*), or synonyms (*mergers* and *takeovers*), their proportion in word type pairs does not seem to give a clear signal regarding the quality of writing.[8]

In contrast, the area of moderately high PMI values (from PMI=2.5 to PMI=3.67 in Figure 2) produces a very consistent picture, with only two points out of 48 in that interval[9] lacking significant positive correlation with essay score (p2 at PMI=3.17 and p5 at PMI=3).

Next, observe the consistent negative correlations between essay score and the proportion of word pairs in bins PMI=0.833 through PMI=1.5. Here again, out of the 30 data points corresponding to these values, only 3 failed to reach statistical significance, although the trend there is still negative.

Finally, there is a trend towards a positive correlation between essay scores and the proportion of mildly negative PMI values (-2<PMI<0), that is, better essays tend to use *more* pairs of disassociated words, although this trend is not as clear-cut as the one on the right-hand side of the distribution.

Assuming that a higher proportion of high PMI pairs corresponds to more topic development and that a higher proportion of negative PMIs corresponds to more creative use of language (in that pairs are chosen that do not generally tend to appear together), it seems that the better essays are *both* more topical *and* more creative than the lower scoring ones. In what follows, we check whether the information about essay quality provided by WAP can be used to improve essay scoring.

[8] It is also possible that some of the instances with very high PMI are pairs that contain low frequency words for which the database predicts a spuriously high PMI based on a single (and a-typical) co-occurrence that happens to repeat in an essay – similar to the *Schwartz eschews* example in (Manning and Schütze, 1999, Table 5.16, p. 181). On the one hand, we do not expect such pairs to occur in any systematic pattern, so they could obscure an otherwise more systematic pattern in the high PMI bins. On the other hand, we do not expect to see many such pairs, simply because a repetition of an a-typical event is likely to be very rare. We thank an anonymous reviewer for suggesting this direction, and leave a more detailed examination of the pairs in the highest-PMI bins to future work.

[9] There are 8 bins of width of 0.167 in the given interval, with 6 datapoints per bin.

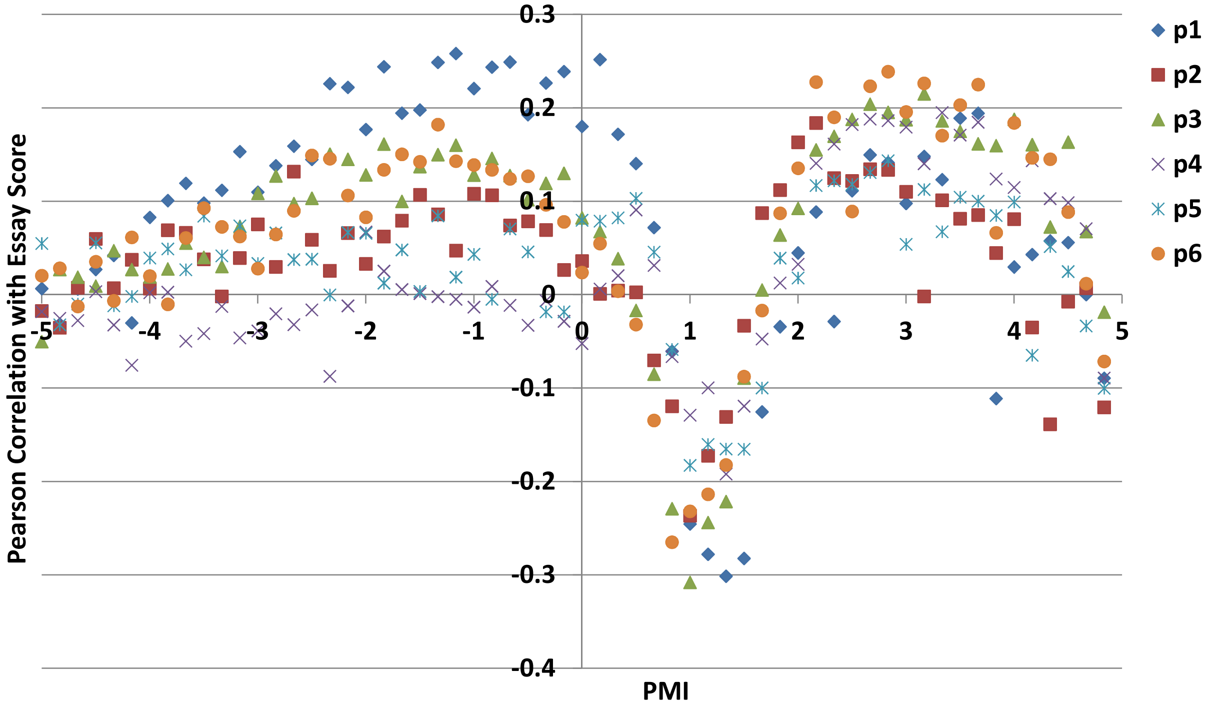

Figure 2: Correlations with essay score for various bins of the WAP histogram. P1 to P6 correspond to the first 6 prompts in SetA.

4.3 Baseline

As a baseline, we use e-rater (Attali and Burstein, 2006), a state-of-art essay scoring system developed at Educational Testing Service.[10] E-rater computes more than 100 micro-features, which are aggregated into macro-features aligned with specific aspects of the writing construct. The system incorporates macro-features measuring grammar, usage, mechanics, style, organization and development, lexical complexity, and vocabulary usage. Table 2 gives examples of micro-features covered by the different macro-features.

E-rater models are built using linear regression on large samples of test-taker essays. We use a generic e-rater model built at Educational Testing Service using essays across a variety of writing prompts, with no connection to the current project and its authors. This model obtains Pearson correlations of r=0.8324-0.8721 with the human scores on setA, and the staggering r=0.9191 and r=0.9146 with the human scores on setB-p1 and setB-p2, respectively. This is a very competitive baseline, as e-rater features explain more than 70% of the variation in essay scores on a relatively coarse scale (setA) and more than 80% of the variation in scores on a fine-grained scale (setB).

Macro-Feature	**Example Micro-Features**
Grammar, Usage, and Mechanics	agreement errors verb formation errors missing punctuation
Style	passive very long or short sentences excessive repetition
Organization and Development	use of discourse elements: thesis, support, conclusion
Lexical Complexity	average word frequency average word length
Vocabulary	similarity to vocabulary in high- vs low-scoring essays

Table 2: Features used in e-rater (Attali and Burstein, 2006).

4.4 Adding WAP

We define **HAT** – high associative tightness – as the percentage of word type pairs with 2.33<PMI≤3.67 (bins PMI=2.5 through PMI=3.67). This range correponds to the longest sequence of adjacent bins in the PMI>0 area that had a positive correlation with essay score in the setA-p1 set. The HAT feature attains significant

[10] http://www.ets.org/erater/about/

(at $p = 0.05$) correlations with essay scores, r=0.11 to r=0.27 for the prompts in setA, and r=0.22 and r=0.21 for the two prompts in setB. We note that the HAT feature is not correlated with essay length. Essay length is not used as a feature in e-rater models, but it typically correlates strongly with the human essay score (at about r=0.70 in our data), as well as with the score provided by e-rater (at about r=0.80).

We also explored a feature that captured the area with the negative correlations identified in section 4.2. This feature did not succeed in improving the performance over the baseline on setA p1-p6; we tentatively conclude that information contained in that feature, i.e. the proprotion of mildly associated vocabulary in an essay, is indirectly captured by another feature or group of features already present in e-rater. Likewise, a feature that calculates the average PMI for all pairs of content word types in the text failed to produce an improvement over the baseline for setA p1-p6. The reason for this can be observed in Figure 2: The higher-scoring essays having more of *both* the low and the high PMI pairs leads to about the same average PMI as for the lower-scoring essays that have a higher concentration of values closer to the average PMI.

4.5 Evaluation

To evaluate the usefulness of WAP in improving automated scoring of essays, we estimate a linear regression model using the human score as a dependent variable (label) and e-rater score and the HAT as the two independent variables (features). The correlations between the two independent variables (e-rater and HAT) are between r=0.11 and r=0.24 on the prompts in setA and setB.

We estimate a regression model on each of setA-pi, $i \in \{1, .., 6\}$, and evaluate them on each of setA-pj, $j \in \{7, .., 9\}$, and compare the performance with that of e-rater alone on setA-pj. Note that e-rater itself is not trained on any of the data in setA and setB; we use the same e-rater model for all evaluations, a generic model that was pretrained on a large number of essays across different prompts. For setB, we estimate the regression model on setB-p1 and test on setB-p2, and vice versa.

Table 3 shows the evaluation results. The HAT feature leads to a statistically significant improve-

Train	Test	E-rater on Test	E-rater+HAT on Test	t
		setA		
p1	p7	0.84043	0.84021	-0.371
p2	p7	0.84043	0.84045	0.408
p3	p7	0.84043	0.83999	-0.597
p4	p7	0.84043	0.84044	0.411
p5	p7	0.84043	0.84028	-0.280
p6	p7	0.84043	0.83926	-1.080
p1	p8	0.83244	0.83316	1.688
p2	p8	0.83244	0.83250	2.234
p3	p8	0.83244	0.83327	1.530
p4	p8	0.83244	0.83250	2.237
p5	p8	0.83244	0.83311	1.752
p6	p8	0.83244	0.83339	1.191
p1	p9	0.86370	0.86612	4.282
p2	p9	0.86370	0.86389	5.205
p3	p9	0.86370	0.86659	4.016
p4	p9	0.86370	0.86388	5.209
p5	p9	0.86370	0.86591	4.390
p6	p9	0.86370	0.86730	3.448
		setB		
p1	p2	0.9146	0.9178	0.983
p2	p1	0.9191	0.9242	2.690

Table 3: Performance of baseline model (e-rater) and models where e-rater was augmented with HAT, a feature based on the word association profile. Performance is measured using Pearson correlation with essay score. We use Wilcoxon Signed-Ranked test for matched pairs, and report the sum of signed ranks (W), the number of ranks (n), and the p value. E-rater+HAT is significantly better than e-rater alone, W=138, n=20, p<0.05. We also measure significance of the improvement for each row individually, using McNemar's test for significance of difference in same-sample correlations (McNemar, 1955, p.148); we report the t value for each test. For values of $t > 1.645$, we can reject the hypothesis that e-rater+HAT is not better than e-rater alone with 95% confidence. Significant improvements are underlined.

ment in the performance of automated scoring. An improvement is observed for 14 out of the 18 evaluations for setA, as well as for both evaluations for setB.[11] Moreover, the largest relative improvement of 0.55%, from 0.9191 to 0.9242, was observed for the setting with the highest baseline performance, suggesting that the HAT feature is still effective even after the delicate ranking of the essays revealed an exceptionally strong performance of e-rater.

5 Related Work

Most of the attention in the computational linguistics research that deals with analysis of the lexis of texts has so far been paid to what in our terms would be the very high end of the word association profile. Thus, following Halliday and Hasan (1976), Hoey (1991), and Morris and Hirst (1991), the notion of *lexical cohesion* has been used to capture repetitions of words and occurrence of words with related meanings in a text. Lexically cohesive words are traced through the text, forming lexical chains or graphs, and these representations are used in a variety of applications, such as segmentation, keyword extraction, summarization, sentiment analysis, temporal indexing, hypelink generation, error correction (Guinaudeau et al., 2012; Marathe and Hirst, 2010; Ercan and Cicekli, 2007; Devitt and Ahmad, 2007; Hirst and Budanitsky, 2005; Inkpen and Désilets, 2005; Gurevych and Strube, 2004; Stokes et al., 2004; Silber and McCoy, 2002; Green, 1998; Al-Halimi and Kazman, 1998; Barzilay and Elhadad, 1997). To our knowledge, lexical cohesion has not so far been used for automated scoring of essays. Our results suggest that this direction is promising, as merely the *proportion* of highly associated word pairs is already contributing a clear signal regarding essay quality; it is possible that additional information can be derived from richer representations common in the lexical cohesion literature.

Aspects related to the distribution of words in essays have been studied in relation to essay scoring. One line of work focuses on assessing coherence of essays. Foltz et al. (1998) use Latent Semantic Analysis to model the smoothness of transitions between adjacent segments of an essay. Higgins et al. (2004) compare sentences from certain discourse segments in an essay to determine their semantic similarity, such as comparing thesis statements to conclusions or thesis statements to essay prompts. Additional approaches include evaluation of coherence based on repeated reference to entities (Burstein et al., 2010; Barzilay and Lapata, 2008; Miltsakaki and Kukich, 2004). Our approach is different in that it does not measure the flow of the text, that is, the sequencing and repetition of the words, but rather assesses the choice of vocabulary as a whole.

Topic models have been proposed as a technique for capturing clusters of related words that tend to occur in the same documents in a given collection. A text is modeled as being composed of a small number of topics, and words in the text are generated conditioned on the selected topics (Gruber et al., 2007; Blei et al., 2003). Since (a) topics encapsulate clusters of highly associated words, and (b) topics for a given text are modeled as being chosen independently from each other, we expect a negative correlation between the number of topics in a document and the tightness of the word association profile of the text.

An alternative representation of word association profile would be a weighted graph, where the weights correspond to pairwise associations between words. Thus, for longer texts, graph analysis techniques would be applicable. Steyvers and Tenenbaum (2005) analyze the graphs induced from large repositories like WordNet or databases of free associations, and find them to be scale-free and small-world; it is an open question whether word association graphs induced from book-length texts would exhibit similar properties.

In the theoretical tradition, our work is closest in spirit to Michael Hoey's theory of lexical priming (Hoey, 2005), positing that users of language internalize patterns of occurrence and non-occurrence of words not only with other words, but also in certain positions in a text, in certain syntactic environments, and in certain evaluative contexts, and use these when creating their own texts. We believe that word association profiles reflect the artwork that goes into using those internalized associations between words when creating a text, achieving the right mix of strong and weak, positive and negative associations.

[11] We also performed a cross-validation test on setA p1-p6, where we estimated a regression model on setA-pi and evaluate it on setA-pj, for all $i, j \in \{1, .., 6\}, i \neq j$, and compared the performance with that of e-rater alone on setA-pj, yielding 30 different train-test combinations. The results were similar to those of the blind test presented here, with e-rater+HAT significantly improving upon e-rater alone, using Wilcoxon test, W=374, n=29, p<0.05.

6 Conclusion

In this paper, we described a new representation of the content vocabulary of a text we call *word association profile* that captures the proportions of highly associated, mildly associated, unassociated, and dis-associated pairs of words selected to co-exist in the given text by its author. We observed that the shape of the distribution is quite stable across various texts, with about half the pairs having a mild association; the allocation of pairs to the higher and the lower levels of association does vary across genres and target audiences.

We further presented a study of the relationship between quality of writing and word association profiles. For a dataset of essays written by college graduates on a number of general topics in a standardized test for graduate school admission and scored by professional raters, we showed that the higher scoring essays tend to have higher percentages of both highly associated and dis-associated pairs, and lower percentagese of mildly associated pairs of words. We hypothesize that this pattern is consistent with the better essays demonstrating both a better topic development (hence the higher percentage of highly related pairs) and a more creative use of language resources, as manifested in a higher percentage of word pairs that generally do not tend to appear together.

Finally, we demonstrated that the information provided by word association profiles leads to a significant improvement in a highly competitive, state-of-art essay scoring system that already measures various aspects of writing quality.

In future work, we intend to investigate in more detail the contribution of various kinds of words to word association profiles, as well as pursue application to evaluation of text complexity.

References

Reem Al-Halimi and Rick Kazman. 1998. Temporal indexing through lexical chaining. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 333–351. Cambridge, MA: MIT Press.

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With e-rater®V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Yigal Attali, Will Lewis, and Michael Steier. 2013. Scoring with the computer: Alternative procedures for improving reliability of holistic essay scoring. *Language Testing*, 30(1):125–141.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of ACL Intelligent Scalable Text Summarization Workshop*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Roberto Basili and Marco Pennacchiotti. 2010. Distributional lexical semantics: Toward uniform representation paradigms for advanced acquisition and processing tasks. *Natural Language Engineering*, 16(4):347–358.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

John Bullinaria and Joseph Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684, Los Angeles, California, June. Association for Computational Linguistics.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic, June. Association for Computational Linguistics.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 334–343, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gonenc Ercan and Ilyas Cicekli. 2007. Using lexical chains for keyword extraction. *Information Processing & Management*, 43(6):1705–1714.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii, October. Association for Computational Linguistics.

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

Michael Flor. 2013. A fast and flexible architecture for very large word n-gram datasets. *Natural Language Engineering*, 19(1):61–93.

Peter Foltz, Walter Kintsch, and Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2):285–307.

Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21(4):353–367.

David Graff and Christopher Cieri. 2003. English Gigaword LDC2003T05. Linguistic Data Consortium, Philadelphia.

Stephen Green. 1998. Automated link generation: Can we do better than term repetition? *Computer Networks*, 30:75–84.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. *Journal of Machine Learning Research - Proceedings Track*, 2:163–170.

Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. 2012. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech and Language*, 26(2):90–104.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of Coling 2004*, pages 764–770, Geneva, Switzerland, August. COLING.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Marti Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 185–192, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press.

Michael Hoey. 2005. *Lexical Priming*. Routledge.

Diana Inkpen and Alain Désilets. 2005. Semantic similarity for detecting recognition errors in automatic speech transcripts. In *Proceedings of Empirical Methods in Natural Language Processing Conference*, pages 49–56, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Michael Jones and Douglas Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1–37.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL*, pages 768–774, Montreal, Canada.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments & Computers*, 28:203–208.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

Meghana Marathe and Graeme Hirst. 2010. Lexical Chains Using Distributional Measures of Concept Distance. In *Proceedings of 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pages 291–302, Iasi, Romania, March.

Quinn McNemar. 1955. *Psychological Statistics*. New York: J. Wiley and Sons, 2nd edition.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappe. 2009. Text segmentation via topic modeling: an analytical study. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1553–1556, New York, NY, USA. ACM.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48.

Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

David Reshef, Yakir Reshef, Hilary Finucane, Sharon Grossman, Gilean McVean, Peter Turnbaugh, Eric Lander, Michael Mitzenmacher, and Pardis Sabeti. 2011. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557, Montréal, Canada, June. Association for Computational Linguistics.

Gerard Salton, Andrew Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Kathy Sheehan, Irene Kostin, and Yoko Futagi. 2008. When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty? In *Proceedings of the Cognitive Science Society*, pages 1978–1983, Washington, DC, July.

Gregory Silber and Kathleen McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29:41–78.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton. 2004. Select: A lexical cohesion based news story segmentation system. *Journal of AI Communications*, 17(1):3–12.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Articial Intelligence Research*, 37:141–188.

Peter D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning*, pages 491–502, Freiburg, Germany, September.

Ziqi Zhang, Anna Gentile, and Fabio Ciravegna. 2012. Recent advances in methods of lexical semantic relatedness – a survey. *Natural Language Engineering*, FirstView:1–69.