Unsupervised Modeling of Topical Relevance in L2 Learner Text

Ronan Cummins
ALTA Institute
Computer Lab
University of Cambridge
UK, CB3 0FD
ronan.cummins@cl.cam.ac.uk

Helen Yannakoudakis
ALTA Institute
Computer Lab
University of Cambridge
UK, CB3 0FD
helen.yannakoudakis@cl.cam.ac.uk

Ted Briscoe
ALTA Institute
Computer Lab
University of Cambridge
UK, CB3 0FD
ejb@cl.cam.ac.uk

Abstract

The automated scoring of second-language (L2) learner text along various writing dimensions is an increasingly active research area. In this paper, we focus on determining the topical relevance of an essay to the prompt that elicited it. Given the burden involved in manually assigning scores for use in training supervised prompt-relevance models, we develop unsupervised models and show that they correlate well with human judgements.

We show that expanding prompts using topically-related words, via pseudo-relevance modelling, is beneficial and outperforms other distributional techniques. Finally, we incorporate our prompt-relevance models into a supervised essay scoring system that predicts a holistic score and show that it improves its performance.

1 Introduction

Given the increase in demand for educational tools and aids for L2 learners of English, the automated scoring of learner texts according to a number of predetermined dimensions (e.g., grammaticality and lexical variety) is an increasingly important research area. While a number of early approaches (Page, 1966; Page, 1994) and recent competitions[1] (Shermis and Hammer, 2012) have sought to assign a holistic score to an entire essay, it is more informative to give detailed feedback to learners by assigning individual scores across each such writing dimension.

This more specific feedback facilitates reflection both on learners' strengths and weaknesses, and focuses attention on the aspects of writing that need improvement. Recent work outlines a number of broad competencies that systems should assess (Kakkonen and Sutinen, 2008). These include *morphology*, *syntax*, *semantics*, *discourse*, and *stylistics*, noting that the specific assessment tasks that might aim to measure these areas of competency may vary. One dimension against which a piece of text is often scored is that of topical relevance. That is, determining if a learner has understood and responded adequately to the prompt. This aspect of automated writing assessment has received considerably less attention than holistic scoring.[2]

Topical relevance is not so much concerned with whether an L2 learner has constructed grammatically correct and well-organised sentences, as it is concerned with whether the learner has understood the prompt and attempted a response with appropriate vocabulary. Other reasons for measuring the topical relevance of a text include the detection of malicious submissions, that is, detecting submissions that have been rote-learned or memorised specifically for assessment situations (Higgins et al., 2006).

In this paper, we employ techniques from the area of distributional semantics and information retrieval (IR) to develop unsupervised prompt-relevance models, and demonstrate that they correlate well with human judgements. In particu-

[1] https://www.kaggle.com/c/asap-aes

[2] We note that a recent paper (Persing and Ng, 2014) has referred to this task as *prompt adherence*, while we use the terms *prompt-relevance* and *topical-relevance* interchangeably throughout this paper.

lar, we study four different methods of expanding a prompt with with topically-related words and show that some are more beneficial than others at overcoming the 'vocabulary mismatch' problem which is typically present in free-text learner writing. To the best of our knowledge, there have been no attempts at a comparative study investigating the effectiveness of such techniques on the automatic prediction of a topical-relevance score in the noisy domain of learner texts, where grammatical errors are common. In addition, we perform an external evaluation to measure the extent to which prompt-relevance *informs* (Rotaru and Litman, 2009) the holistic score.

The remainder of the paper is outlined as follows: Section 2 discusses related work and outlines our contribution. Section 3 presents our framework and four unsupervised approaches to measuring semantic similarity. Section 4 presents both quantitative and qualitative evaluations for all of the methods employed in this paper. Section 5 performs an external evaluation by incorporating the best prompt-relevance model as features into a supervised preference ranking approach. Finally, Section 6 concludes with a discussion and outline of future work.

2 Related Research

There are a number of existing automated text-scoring systems (sometimes referred to as *essay scoring systems*). For an overview, the interested reader is directed to reviews and advances in the area (Shermis and Burstein, 2003; Landauer, 2003; Valenti et al., 2003; Dikli, 2006; Phillips, 2007; Briscoe et al., 2010; Shermis and Burstein, 2013). In this section, we review related research on topical-relevance detection for automated writing assessment, and outline the key differences between our approach and that of existing work.

A wide variety of computational approaches (Miller, 2003; Landauer et al., 2003; Higgins et al., 2004; Higgins and Burstein, 2007; Chen et al., 2010) have been used to automatically assess L2 texts. Early work on topical relevance (Higgins et al., 2006) posed the problem as one of binary classification and aimed to identify whether a text was either on or off-topic. The main motivation of the research was to detect off-topic text, text submitted mistakenly (within an online assessment setting), or text submitted in bad faith (i.e., possibly memorised on an unrelated topic). They adopted an unsupervised approach to the problem, where they matched each text to its corresponding prompt using tf-idf weighted content vectors and a similarity function. One of the heuristic approaches employed in that work was to calculate the similarity of an essay to a number of unrelated prompts. If the essay was closer to an unrelated prompt than the relevant one, the essay was deemed to be off-topic.

Briscoe et al. (2010) tackle the problem of off-topic detection using more complex distributional semantic models that tend to overcome the problem of vocabulary mismatch. However, they frame the task as binary classification and evaluate their approach by determining if it can associate a learner text with the correct prompt. The work which is closest in spirit to that of our own is by Louis and Higgins (2010), who expand prompts using morphological variations, synonyms, and words that are distributionally similar to those that appear in the prompt. Their work builds on the earlier work by Higgins et al. (2006), and again pose the problem as one of binary classification.

The most recent work of Persing and Ng (2014) involves scoring L2 learner texts for relevance on a seven-point scale using a feature-rich linear regression approach. While they demonstrate that learning one linear regression model per prompt is a useful supervised approach, it means that substantial training data is needed for each prompt in order to build the models. For the task of determining topical relevance, this places a substantial burden on manually annotating texts for each individual prompt.[3] As a result, supervised prompt-specific approaches are impractical and less flexible in an operational setting; if, for example, a new previously-unseen prompt is required for an upcoming assessment, the model cannot be applied until a sizeable amount of manually-annotated response texts are collected and annotated for that prompt.

A dataset developed from the international corpus of learner data (ICLE) (Granger et al., 2009) consisting of 830 essays measured for relevance against one of 13 prompts on a seven-point scale was re-

[3] In fact, it is often the case that there are multiple prompts per exam, which change for every exam sitting.

leased as part of that work (Persing and Ng, 2014). We make use of this new resource in our work as it is the *only* such public dataset.[4] We make the following contributions to the automated assessment of topical relevance:

- We perform the first systematic comparison of several unsupervised methods for assessing topical relevance in L2 learner text on a publicly available dataset.
- We adopt a new unsupervised pseudo-relevance feedback language-modelling approach and show that it correlates well with human judgements and outperforms a number of other distributional approaches.
- We perform an external evaluation of our best prompt-relevance models by incorporating them into the feature set of a supervised prompt-independent text-scoring system, and show that they improve its performance.

3 Semantic Prompt Relevance

Previous research (Higgins et al., 2006) has shown that representing a prompt p and an essay s as tf-idf weighted vectors[5] $\boldsymbol{p}$ and $\boldsymbol{s}$ in the term space $\mathbb{R}^v$ (where v is the vocabulary of the system) yields useful representations for exact matching using cosine similarity as follows:

$$cos(\boldsymbol{p}, \boldsymbol{s}) = \frac{\sum_{t \in v} p_t \cdot s_t}{\sqrt{\sum_{t \in v} p_t^2 \cdot \sum_{t \in v} s_t^2}} \qquad (1)$$

However, it is likely that many L2-learner texts will use words that are related to the prompt, but which do not have an exact match to any words contained in the prompt. Therefore, we extend this approach by aiming to expand the prompt p with a set of topically related expansion terms e using one of a number of distributional similarity techniques.

3.1 Prompt Expansion

As a general method of prompt expansion, we represent the prompt p and each candidate expansion word w as vectors $\boldsymbol{p}$ and $\boldsymbol{w}$ in an n-dimensional space $\mathbb{R}^n$, and then use some measure of similarity between the two vectors (e.g. cosine similarity) to rank the candidate expansion words according to how close they are to the original prompt. We then select the top $|e|$ most similar expansion terms to add to the original prompt.

Once the $|e|$ closest terms are selected and added to the original prompt p, we create a tf-idf weighted expanded prompt vector $\boldsymbol{p}_{p+e}$ and compare it to the tf-idf essay vector $\boldsymbol{s}$ using cosine similarity in the original space $\mathbb{R}^v$ as per Equation (1). In our approach, we conduct the essay matching in the term space $\mathbb{R}^v$ as it allows us to analyse the quality of the expansion terms, and subsequently to understand the merits and demerits of the various approaches. We now outline four methods of selecting candidate prompt expansion terms.

3.2 Traditional Distributional Semantics

Our first approach involves building traditional distributional vectors by constructing a matrix of co-occurrence frequencies. For a specific word w, its vector is constructed by counting the words (its context words c) that it co-occurs with in a specified context (usually a window of a few words). The row for a specific word w then represents the vector for that word. We weight the vector elements using the PPMI (positive pointwise mutual information) weighting scheme (Turney et al., 2010).

We build word vectors using a lemmatised version of Wikipedia from 2013. We removed from the corpus all words that appeared less than 200 times and used the 96,811 remaining words as both potential expansion words w and as contexts c. We used a 5 word context window (2 words either side of the target word) and reduced the size of the resultant vectors by only storing dimensions that had a PPMI greater than 2.0 (Turney et al., 2010). The resultant vectors are competitive with the best reported results for traditional word vectors on a word–word similarity task (Spearman-$\rho = 0.732$ on 3000 word-pairs from the MEN dataset) (Levy et al., 2015). We create a vector representation for the prompt $\boldsymbol{p}$ in $\mathbb{R}^n$ by summing the PPMI word-vectors of the words occurring in the prompt. Finally, the $|e|$ closest words to the prompt vector $\mathbf{p}$, as measured by cosine similarity, can then be selected as expansion terms.

[4] www.hlt.utdallas.edu/~persingq/ICLE/paDataset.html

[5] We use bold lower-case letters throughout to denote vectors, including probability vectors.

3.3 Random Indexing

Random Indexing (RI) (Kanerva et al., 2000) is an approach which incrementally builds word vectors in a dimensionally-reduced space. Words are initially assigned a unique random index vector in a space $\mathbb{Z}^n$, where n is user-defined. These near-orthogonal vectors are updated by iterating over a corpus of text. In particular, the word vector for a specific word w is altered by adding to it the vectors of the words in its contexts. The process proceeds incrementally and therefore only requires one pass over the data. In this way, words that occur in similar contexts will be pushed towards similar points in the space $\mathbb{Z}^n$.

We use *Random Indexing* to build word vectors using the S-Space package[6] using the same preprocessed Wikipedia corpus as outlined in the previous section. We used a dimensionality of 400 with window sizes up to 5 words (finding a window of 5 words to create better vectors for the word-word similarity task). The resultant vectors are not as competitive as those built using the traditional approach on a word-word similarity task (Spearman-$\rho = 0.432$ on 3000 word-pairs from the MEN dataset). Again, we create a vector representation for the prompt $\boldsymbol{p}$ by summing the RI vectors, and find the closest words vectors $\boldsymbol{w}$ to the prompt.

3.4 Word Embeddings

The *continuous bag-of-words* architecture (cbow) and the *skip-gram* architectures (skip) in **word2vec** have been shown to be particularly well-suited to learning *word-embeddings* (i.e. low-dimensional vector representations of words) (Mikolov et al., 2013). The **word2vec** package[7] from Mikolov is the original implementation of these models.

We use **word2vec** to learn distributed representations for prompts in a similar manner to that just outlined (in Section 3.2 and Section 3.3). In particular, we learn distributed vectors using both *cbow* and *skip* and the same preprocessed version of Wikipedia as used previously. We used word vectors of length 400 for both architectures with a window of 5 for *cbow* and 10 for *skip-gram* as recommended in the original documentation. For both approaches we use negative sampling. The performance of these approaches on the word-word MEN dataset are $\rho = 0.737$ and $\rho = 0.764$ for *cbow* and *skip* respectively. As with previous approaches, we create a vector representation for the prompt $\boldsymbol{p}$ by summing the vectors of the words in the prompt.

[6] https://github.com/fozziethebeat/S-Space

[7] `https://code.google.com/p/word2vec/`

3.5 Pseudo-Relevance Feedback

Pseudo-relevance feedback (PRF) is a technique in IR for expanding queries with topically related words. In PRF, the top $|F|$ ranked documents for a query are deemed relevant and candidate terms occurring in these documents are analysed and selected according to a term-selection function. Each candidate word can be viewed as being described by a vector of contexts of dimensionality $|F|$ (i.e. where the entire document $d \in F$ is the context).

We use this approach by using a prompt analogously to a query. In the popular relevance modelling (RM) framework (Lv and Zhai, 2009), the term-selection value can be viewed as the dot-product of a prompt vector $\boldsymbol{p}$ (the vector of similarities between the initial prompt p and the document contexts $d \in F$) and the candidate word vector $\boldsymbol{w}$ (the vector of weights for word w in its contexts $d \in F$) as follows:

$$PRF(\boldsymbol{p}, \boldsymbol{w}) = \sum_{d \in F} f(w, d) \cdot Pr(p|d) \qquad (2)$$

where $f(w, d)$ is the weight of the candidate word w in document d and $Pr(p|d)$ is the probability that d generated p, i.e. the query-likelihood (Ponte and Croft, 1998). Furthermore, by selecting only the most important dimensions (i.e. top $|F|$ documents), dimensional reduction is automatically incorporated in an operationally efficient manner. PRF can be viewed as a dimensionally-reduced probabilistic version of Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007). The typical dimensionality used for PRF is usually of around $|F| = 20$.

In the language modelling framework, documents are assumed to have been generated by a mixture of a topical model $\boldsymbol{\alpha}_\tau$ and a background model $\boldsymbol{\alpha}_c$, such that $d \sim (1 - \omega) \cdot \boldsymbol{\alpha}_\tau + \omega \cdot \boldsymbol{\alpha}_c$ where ω is the mixture parameter. Given a candidate term w

appearing in d, the probability that it was generated by the topical model is as follows:

$$f(w,d) = p(\boldsymbol{\alpha}_\tau|w) = \frac{(1-\omega)\cdot\boldsymbol{\alpha}_\tau}{(1-\omega)\cdot\boldsymbol{\alpha}_\tau + \omega\cdot\boldsymbol{\alpha}_c} \quad (3)$$

and therefore, we use this probability of topicality $f(w,d)$ as the vector weights for $\boldsymbol{w}$. Assuming that documents have been generated by a multivariate Pólya distribution (Cummins et al., 2015), $f(w,d)$ is as follows:

$$f(w,d) = \frac{tf_{w,d}}{tf_{w,d} + \frac{\omega\cdot m_c\cdot df_w}{(1-\omega)\cdot\sum_{w'} df_{w'}} \cdot \frac{|d|}{m_d}} \quad (4)$$

where $tf_{w,d}$ is the term-frequency, df_w is the document frequency of w in the collection being searched, m_d is the number of unique terms in the document, m_c is the background mass (Cummins et al., 2015), and $\omega = 0.8$ is a stable hyper-parameter that controls the belief in the background model. Essentially, this approach (denoted PRF) selects terms that occur more frequently in the top $|F|$ documents than they should by chance. As our documents, we use the same preprocessed Wikipedia corpus as outlined previously.

4 Evaluation of Expansion Methods

In this section, we present results on the effectiveness of the unsupervised approaches for the task of assessing the prompt relevance of an essay.

4.1 Data and Experimental Setup

For the first set of experiments, we use 830 L2 learner essays from the ICLE dataset that are assessed for prompt relevance across 13 prompts. This corpus consists of essays written by higher intermediate to advanced learners of English, which corresponds to approximately B2 level, or above, of the CEFR (Common European Framework of Reference for Languages). The scores assigned to the essays range from 1.0 to 4.0 in increments of 0.5 (although all essays received a score of 2.0 or more in the dataset as seen in Table 1). The essays were double-marked and the linear correlation[8] between the assessors was 0.243 (a weak correlation). The distribution of essays per prompt is included in Table 2. We lemmatised all prompts and essays using RASP (Briscoe et al., 2006). A point worth noting is that there are minimal essay-length effects in operation on this dataset. The Spearman correlation between the length of the essay and the human-assigned prompt-relevance score across all 830 essays is $\rho = 0.007$.

score	1.0	1.5	2.0	2.5	3.0	3.5	4.0
# of essays	0	0	8	44	105	230	443

Table 1: Distribution of ICLE essays over score grades.

As a baseline approach, we use the cosine similarity between the original prompt (unexpanded) and the essay $cos(p,s)$. For all expansion approaches, we set the number of expansion terms $|e| = 200$ and use the weight of association between the prompt and the expansion term as the expansion term's frequency tf value in the expanded prompt. We evaluate the approaches by calculating Spearman's rank (ρ) correlation coefficient between each method's predicted similarity score and the scores assigned by the assessors.

4.2 Results for Prompt Relevance

Table 2 (Top) shows the performance of the approaches over 11 prompts.[9] On average, all approaches increase over the baseline. We can see that the most consistent approach is the PRF approach as it improves over the baseline in 10 out of 11 prompts. The RI approach also performs well and is the best approach on many of the prompts.

However, to measure the topical quality of the expansion words selected by each approach in isolation, we removed the original prompt words from the expanded prompts and again calculated the performance of the different approaches. This more rigorous evaluation in Table 2 (Bottom) shows that the topical quality of the expansion words from the PRF approach tends to be better than the other approaches. We next look at the actual expansion words selected for two prompts.

[8] While this seems to suggest that the upper-bound on this dataset is quite low, the original work notes that 89% of the time, assessors graded within a point of each other. Furthermore, correlation is affected by scale (Yannakoudakis and Cummins, 2015).

[9] The two remaining prompts have only three essays associated with them.

Prompt	1	2	3	4	5	6	7	8	9	10	11	Mean
# of essays	237	53	64	58	131	43	80	28	49	71	13	
length	-0.113	-0.026	-0.062	0.211	-0.023	-0.111	0.103	-0.115	-0.056	**0.171**	0.520	0.045
$cos(p, s)$	0.324	0.120	**0.195**	0.122	0.205	-0.019	0.333	0.511	0.268	0.064	0.637	0.251
ds_{p+e}	0.328	0.141	0.182	0.114	0.208	-0.011	0.340	0.519	0.280	0.082	0.637	0.256
RI_{p+e}	**0.372**	0.098	0.103	**0.214**	0.192	**0.093**	**0.398**	**0.720**	0.259	0.116	0.449	0.274
$cbow_{p+e}$	0.345	0.125	0.131	0.114	0.209	0.068	0.328	0.581	0.265	-0.024	0.637	0.253
$skip_{p+e}$	0.359	0.160	0.183	0.139	0.245	0.026	0.363	0.571	0.278	-0.064	0.677	0.267
PRF_{p+e}	0.348	**0.188**	0.126	0.145	**0.260**	0.034	0.340	0.598	**0.335**	0.078	**0.679**	**0.285**

Prompt	1	2	3	4	5	6	7	8	9	10	11	Mean
# of essays	237	53	64	58	131	43	80	28	49	71	13	
$ds_{e\setminus p}$	0.008	0.043	-0.098	-0.073	-0.017	-0.092	0.126	0.619	**0.202**	0.029	0.375	0.102
$RI_{e\setminus p}$	**0.097**	0.016	-0.195	0.326	0.061	0.091	**0.206**	0.572	0.030	**0.185**	-0.082	0.119
$cbow_{e\setminus p}$	0.080	0.025	-0.209	0.165	0.071	**0.266**	0.088	**0.677**	-0.079	-0.118	0.239	0.110
$skip_{e\setminus p}$	0.087	0.133	**-0.052**	0.167	0.149	0.188	0.173	0.592	0.000	-0.171	0.222	0.135
$PRF_{e\setminus p}$	0.079	**0.184**	-0.055	**0.363**	**0.151**	0.155	0.157	0.612	0.161	0.125	**0.455**	**0.217**

Table 2: Correlation (Spearman's ρ) between prompt–essay similarity scores and human annotations for each prompt (higher values indicate a better approach) for expansion methods when including original prompt terms (Top – denoted $p + e$) and when removing original prompt terms from the expanded prompt (Bottom – denoted $e \setminus p$). Best result in bold.

# 2 – Most University degrees are theoretical and do not prepare us for the real life. Do you agree or disagree?									
university degree theoretical prepare real life									
ds		RI		cbow		skip		PRF	
faculty	0.222	accept	0.948	accept	0.598	however	0.677	theory	0.544
graduate	0.214	while	0.919	psychology	0.558	nevertheless	0.677	study	0.444
professor	0.21	experience	0.918	understand	0.553	indeed	0.675	science	0.414
phd	0.204	idea	0.915	study	0.551	insist	0.672	differ	0.396
mathematics	0.199	from	0.913	teach	0.55	accept	0.671	student	0.396
philosophy	0.195	work	0.912	philosophy	0.549	fact	0.67	philosophy	0.394
theory	0.194	acknowledge	0.911	knowledge	0.545	s	0.67	topic	0.392
sociology	0.189	nevertheless	0.911	argument	0.538	would	0.664	educate	0.372
science	0.185	notice	0.91	discuss	0.538	while	0.66	academy	0.361
study	0.182	nonetheless	0.909	theory	0.528	nonetheless	0.656	argue	0.354

# 9 – Feminists have done more harm to the cause of women than good.									
feminist harm cause women									
ds		RI		cbow		skip		PRF	
symptom	0.310	likewise	0.883	feminism	0.612	feminism	0.671	feminism	0.910
disease	0.275	consequence	0.882	sexual	0.583	landdyke	0.632	sex	0.896
risk	0.270	furthermore	0.879	violence	0.577	woman	0.617	sexual	0.896
chronic	0.266	affect	0.875	stigmata	0.573	affect	0.594	oppress	0.883
treatment	0.260	response	0.873	perceive	0.573	twwa	0.580	argument	0.875
infect	0.256	moreover	0.871	affect	0.564	argue	0.580	rape	0.800
diagnosis	0.255	hinder	0.871	detriment	0.553	provoke	0.578	men	0.787
patient	0.255	expose	0.869	homosexual	0.547	believe	0.574	gender	0.762
induce	0.253	lastly	0.866	consequence	0.545	consequence	0.573	anti	0.749
disorder	0.247	perceive	0.863	oppress	0.545	sexism	0.573	right	0.740

Table 3: The top 10 non-prompt words and their similarity to the prompt in a lemmatised Wikipedia corpus of 4.4M documents.

4.3 Qualitative Evaluation of Expansion Terms

Table 3 shows the expansion words selected by each approach for two prompts (prompts # 2 and # 9). For prompt # 2 we can see the top words selected for RI and *skip* do not seem topically similar to the prompt. The top words for *ds*, *cbow*, and PRF seem on-topic and might be part of useful feedback to a learner writing for this prompt.

For prompt # 9, *ds* and RI do not tend to promote topically related words. The words for the *ds* approach seem to be related to topic of *diseases* as it may have been mislead by some of the prompt words. In fact, the top terms promoted by the RI approach are not particularly on-topic for any of the 11 prompts, despite the empirical evaluation in the previous section. This could be because some topical words appear further down the ranking for RI.

We believe the main reason that the PRF approach outperforms the others is that topicality is a quality that spans larger segments of text (e.g. docu-

ments). For the other approaches, the words that are promoted are very close in proximity to the prompt words (due to the smaller context sizes), and this is more likely to capture local aspects of word usage. Furthermore, in the PRF approach the most important contexts are those in which *all* prompt words appear together, and this aids automatic disambiguation. Regardless, due to the empirical results in the previous section and the perceived topical quality of the terms from the PRF approach, we make use of the PRF approach as a feature in the next experiment.

5 Prompt-Relevance for Holistic Scoring

We now evaluate the effectiveness of a supervised essay scoring system that incorporates tf-idf similarity features and the PRF approach for the task of predicting an overall essay quality score.

5.1 Data and Experimental Setup

For this experiment, we used a dataset consisting of 2,316 essays written for the IELTS (International English Language Testing System) English examination from 2005 to 2010 (Nicholls, 2003). The examination is designed to measure a broad proficiency continuum ranging from an intermediate to a proficient level of English (A2 to C2 in the CEFR levels). The essays are associated with 22 prompts that are similar in style (i.e. essay style) to those in the ICLE dataset. Candidates are assigned an overall score on a scale from 1 to 9. Prompt relevance is an aspect that is present in the marking criteria, and it is identified as a determinant of the overall score. We therefore hypothesise that adding prompt-relevance measures to the feature set of a prompt-independent essay scoring system (i.e. that is designed to assess linguistic competence only) would better reflect the evaluation performed by examiners and improve system performance.

The baseline system is a linear preference ranking model (Yannakoudakis et al., 2011; Yannakoudakis and Briscoe, 2012) and is trained to predict an overall essay score based on the following set of features:

- word unigrams, bigrams, and trigrams
- POS (part-of-speech) counts
- grammatical relations
- essay length (# of unique words)
- counts of cohesive devices
- max-word length and min-sentence length
- number of errors based on a presence/absence trigram language model

We divided the dataset into 5-folds in two separate ways. First, we created *prompt-dependent* folds, where essays associated with all 22 prompts appear in both the training and test data in the appropriate proportions. This scenario allows the system to learn from essays that were written in response to the prompt. Second, we created *prompt-independent* folds, where all essays associated with a specific prompt appear in only one fold. This second dataset is a more realistic real-world scenario (see Section 2) whereby the system learns on one set of prompts (possibly from previous years) and aims to predict the score for essays associated with different prompts. For both of these supervised experiments, we measured system performance using Spearman's and Pearson's correlation between the output of the system and the gold essay scores (human judgements).

In order to examine the effect of prompt relevance on these datasets, we added to our baseline system two sets of features. The first set of features labelled PR includes the cosine similarity between the essay and the prompt $cos(p, s)$, the fraction of essays words that appear in the prompt $cov(p, s)$, and the fraction of prompt words that appear in the essay $cov(s, p)$. The second set of features labelled semPR is the same as the first set except that the prompt is expanded using the PRF method from earlier.

5.2 Results for Overall Scoring

The results of the experiment are outlined in Table 4. Firstly, we observe that the effectiveness of the baseline system is higher on the *prompt-dependent* folds ($\rho = 0.661$) than on the *prompt-independent* folds ($\rho = 0.637$). This confirms expectations as the *prompt-dependent* folds allow the baseline model to learn useful features from essays written specifically for those prompts. When adding the exact matching prompt-relevance features – referred to as PR in Table 4 – we observe an increase in performance on the *prompt-independent* folds. When we add the semantic prompt-relevance models – referred to as semPR in Table 4 – we again observe a modest increase in

Prompt-Dependent Folds		
System	Spearman-ρ	Pearson-r
Baseline	0.661	0.686
+ PR	0.659	0.685
+ semPR	0.662	0.691

Prompt-Independent Folds		
System	Spearman-ρ	Pearson-r
Baseline	0.637	0.665
+ PR	0.650†	0.678†
+ semPR	0.656†	0.687†

Table 4: Performance of systems using 5-fold cross-validation on prompt-dependent folds (top) and prompt-independent folds (bottom) when adding unsupervised prompt-relevance (PR) features and semantic prompt-relevance features (semPR) on a set of 2316 essays. † means statistically significant compared to the baseline using Steiger's test (1980).

performance on the *prompt-independent* folds. We can see that both Spearman and Pearson correlations approach the performance of the baseline system on the *prompt-dependent* folds.

On the other hand, there is little or no increase in performance when adding the PR and semPR features on the *prompt-dependent* folds. One suspected reason for this is that it is likely that the lexical features in the *prompt-dependent* folds are performing prompt-relevance modelling (by learning appropriate weights for lexical features in essays written for that prompt). Overall, this is an interesting result as it shows that the features developed in this paper are useful and contribute to the holistic score in real-world examinations.

6 Discussion

Firstly, the results from Section 4 are not directly comparable with previous research using the ICLE dataset, as that work (Persing and Ng, 2014) reported metrics averaged over all essays where each prompt was not isolated individually. Ignoring prompt effects may lead to favouring systems that perform well only on a few prompts, and that are not robust across the types of prompt that may be used operationally. Table 5 shows the results of the approaches outlined in this paper against those from the original research using the ICLE dataset that used supervised models. Importantly, we achieve these correlations without any training data.

System	Baseline*	tf-idf	PRF	Persing*
Pearson's-r	0.233	0.261	0.277	0.360

Table 5: Pearson correlation of systems over all 830 essays. * means from original paper.

Interestingly, we have shown that the PRF prompt expansion is effective and is easily analysable. In an operational setting, prompt expansion is likely to be a highly important feature. Observing non-prompt words, that are related to the prompt, in a learner text is likely to be indicative of a learner who has a good understanding of the vocabulary of the topic.

The expansion step issues the entire prompt to a Wikipedia index to gather candidate expansion terms. While this has been shown to be a useful approach on average, there may be cases when aspects of the prompt are not adequately reflected by the candidate expansion terms. In such cases it may be better to *partition* the prompt into useful phrases that can be expanded in isolation, or to manually rephrase the prompt before expanding it with related terms.

6.1 Conclusion and Future Work

We have shown that using an unsupervised pseudo-relevance language modelling approach to measuring relevance in learner texts is beneficial as it correlates with human annotators. The expansion terms in isolation have been shown to be useful and we argue that they are an important feature for overcoming vocabulary mismatch in learner text.

The estimation of an L2 learner's language model from lexemes produced by the learner is an intuitive and theoretically-motivated way to assess many lexical aspects of writing. However, compositionally-motivated language modelling approaches exist (Mitchell and Lapata, 2009), and it would be interesting to investigate these across different areas in assessment.

The approaches developed herein may also be useful for providing feedback and/or suggestions to learners during the process of writing. Future work will look at supplying feedback in pedagogically sound ways.

Acknowledgements

We would like to thank Cambridge English Language Assessment for supporting this research, and the anonymous reviewers for their useful feedback.

References

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80. ACL.

Ted Briscoe, Ben Medlock, and Øistein Andersen. 2010. Automated assessment of esol free text examinations. *University of Cambridge Computer Laboratory Technical Reports*, UCAM-CL-TR-790.

Y. Y. Chen, C. L. Liu, T. H. Chang, and C. H. Lee. 2010. An Unsupervised Automated Essay Scoring System. *IEEE Intelligent Systems*, 25(5):61–67.

Ronan Cummins, Jiaul H. Paik, and Yuanhua Lv. 2015. A Pólya urn document language model for improved information retrieval. *ACM Transactions of Informations Systems*, 33(4):21.

Semire Dikli. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1).

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. The international corpus of learner english. version 2. handbook and cd-rom. *Louvain- la- Neuve:Presses universitaires de Louvain.*

Derrick Higgins and Jill Burstein. 2007. Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 185–192. ACL.

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

Tuomo Kakkonen and Erkki Sutinen. 2008. Evaluation criteria for automatic essay assessment systems–there is much more to it than just the correlation. In *International Conference on Computers in Education (ICCE)*, pages 111–116.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–6. Erlbaum.

Thomas K Landauer, Darrell Laham, and Peter W Foltz. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

Thomas K Landauer. 2003. Automatic essay assessment. *Assessment in education: Principles, policy & practice*, 10(3):295–308.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

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. ACL.

Yuanhua Lv and ChengXiang Zhai. 2009. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1895–1898, New York, NY, USA. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tristan Miller. 2003. Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, 29(4):495 – 512.

Jeff Mitchell and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 430–439. ACL.

Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *In Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581.

Ellis B Page. 1966. The imminence of grading essays by computer. *Phi Delta Kappan*, 47:238–243.

Ellis Batten Page. 1994. Computer grading of student prose, using modern concepts and software. *The Journal of experimental education*, 62(2):127–142.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1534–1543, Baltimore, Maryland, June. ACL.

Susan Phillips. 2007. *Automated essay scoring: A literature review*, volume 30. Society for the Advancement of Excellence in Education (SAEE).

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Mihai Rotaru and Diane J Litman. 2009. Discourse structure and performance analysis: Beyond the correlation. In *Proceedings of the SIGDIAL 2009 Conference: The 10th annual meeting of the special interest group on discourse and dialogue*, pages 178–187. Association for Computational Linguistics.

Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

M Shermis and B Hammer. 2012. Contrasting state-of-the-art automated scoring of essays: analysis. Technical report, The University of Akron and Kaggle.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2:3–118.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in esol learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43. Association for Computational Linguistics.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.