Off-topic essay detection using short prompt texts

Annie Louis
University of Pennsylvania
Philadelphia, PA 19104, USA
lannie@seas.upenn.edu

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

Abstract

Our work addresses the problem of predicting whether an essay is off-topic to a given prompt or question *without* any previously-seen essays as training data. Prior work has used similarity between essay vocabulary and prompt words to estimate the degree of on-topic content. In our corpus of opinion essays, prompts are very short, and using similarity with such prompts to detect off-topic essays yields error rates of about 10%. We propose two methods to enable better comparison of prompt and essay text. We automatically expand short prompts before comparison, with words likely to appear in an essay to that prompt. We also apply spelling correction to the essay texts. Both methods reduce the error rates during off-topic essay detection and turn out to be complementary, leading to even better performance when used in unison.

1 Introduction

It is important to limit the opportunity to submit uncooperative responses to educational software (Baker et al., 2009). We address the task of detecting essays that are irrelevant to a given prompt (essay question) when training data is *not* available and the prompt text is *very short*.

When example essays for a prompt are available, they can be used to learn word patterns to distinguish on-topic from off-topic essays. Alternatively, prior work (Higgins et al., 2006) has motivated using similarity between essay and prompt vocabularies to detect off-topic essays. In Section 2, we examine the performance of prompt-essay comparison for four different essay types. We show that in the case of prompts with 9 or 13 content words on average, the error rates are higher compared to those with 60 or more content words. In addition, more errors are observed when the method is used on essays written by English language learners compared to more advanced test takers. An example short prompt from our opinion essays' corpus is shown below. Test-takers provided arguments for/or against the opinion expressed by the prompt.

[1] "In the past, people were more friendly than they are today."

To address this problem, we propose two enhancements. We use unsupervised methods to expand the prompt text with words likely to appear in essays to that prompt. Our approach is based on the intuition that regularities exist in the words which appear in essays, beyond the prevalence of actual prompt words. In a similar vein, misspellings in the essays, particulary of the prompt words, are also problematic for prompt-based methods. Therefore we apply spelling correction to the essay text before comparison. Our results show that both methods lower the error rates. The relative performance of the two methods varies depending on the essay type; however, their combination gives the overall best results regardless of essay type.

2 Effect of prompt and essay properties

In this section, we analyze the off-topic essay prediction accuracies resulting from direct comparison of original prompt and essay texts. We use four different corpora of essays collected and scored during high stakes tests with an English writing component. They differ in task type and average prompt length, as well as the skill level expected from the test taker.

In one of the tasks, the test taker reads a passage and listens to a lecture and then writes a summary of the main points. For such essays, the prompt text (reading passage plus lecture transcript) available for comparison is quite long (about 276 content words). In the other 3 tasks, the test taker has to provide an argument for or against some opinion expressed in the prompt. One of these has long prompts (60 content words). The other two involve only single sentence prompts as in example [1] and have 13 and 9 content words on average. Two of these tasks focused on English language learners and the other two involved advanced users (applicants to graduate study programs in the U.S.). See Table 1 for a summary of the essay types.[1]

2.1 Data

For each of the task types described above, our corpus contains essays written to 10 different prompts. We used essays to 3 prompts as development data. To build an evaluation test set, we randomly sampled 350 essays for each of the 7 remaining prompts to use as positive examples. It is difficult to assemble a sufficient number of naturally-occurring off-topic essays for testing. However, an essay on-topic to a particular prompt can be considered as *pseudo off-topic* to a different prompt. Hence, to complement the positive examples for each prompt, an equal number of negative examples were chosen at random from essays to the remaining 6 prompts.

2.2 Experimental setup

We use the approach for off-topic essay detection suggested in prior work by Higgins et al. (2006). The method uses cosine overlap between tf*idf vectors of prompt and essay content words to measure the similarity between a prompt-essay pair.

$$sim(prompt, essay) = \frac{v_{essay} . v_{prompt}}{\|v_{essay}\| \|v_{prompt}\|} \quad (1)$$

An essay is compared with the *target* prompt (prompt with which topicality must be checked) together with a set of *reference* prompts, different from the target. The reference prompts are also chosen to be different from the actual prompts of the negative examples in our dataset. If the target prompt is ranked as *most similar*[2] in the list of compared prompts, the essay is classified as on-topic. 9 reference prompts were used in our experiments.

Type	Skill	Prompt len.	Avg FP	Avg FN
1	Learners	276	0.73	11.79
2	Advanced	60	0.20	6.20
3	Advanced	13	2.94	8.90
4	Learners	9	9.73	11.07

Table 1: Effect of essay types: average prompt length, false positive and false negative rates

We compute two error rates.

FALSE POSITIVE - percentage of on-topic essays incorrectly flagged as off-topic.

FALSE NEGATIVE - percentage of off-topic essays which the system failed to flag.

In this task, it is of utmost importance to maintain very low false positive rates, as incorrect labeling of an on-topic essay as off-topic is undesirable.

2.3 Observations

In Table 1, we report the average false positive and false negative rates for the 7 prompts in the test set for each essay type. For long prompts, both *Types 1* and *2*, the false positive rates are very low. The classification of *Type 2* essays which were also written by advanced test takers is the most accurate.

However, for essays with shorter prompts (*Types 3 and 4*), the false positive rates are higher. In fact, in the case of *Type 4* essays written by English language learners, the false positive rates are as high as 10%. Therefore we focus on improving the results in these two cases which involve short prompts.

Both prompt length and the English proficiency of the test taker seem to influence the prediction accuracies for off-topic essay detection. In our work, we address these two challenges by: a) automatic expansion of short prompts (Section 3) and b) correction of spelling errors in essay texts (Section 4).

3 Prompt expansion

We designed four automatic methods to add relevant words to the prompt text.

[1]Essay sources: Type 1-TOEFL integrated writing task, Type 4-TOEFL independent writing task, Types 2 & 3-argument and issue tasks in Analytical Writing section of GRE

[2]Less strict cutoffs may be used, for example, on-topic if target prompt is within rank 3 or 5, etc. However even a cutoff of 2 incorrectly classifies 25% of off-topic essays as on-topic.

3.1 Unsupervised methods

Inflected forms: Given a prompt word, *"friendly"*, its morphological variants—*"friend"*, *"friendlier"*, *"friendliness"*—are also likely to be used in essays to that prompt. Inflected forms are the simplest and most restrictive class in our set of expansions. They were obtained by a rule-based approach (Leacock and Chodorow, 2003) which adds/modifies prefixes and suffixes of words to obtain inflected forms. These rules were adapted from WordNet rules designed to get the base forms of inflected words.

Synonyms: Words with the same meaning as prompt words might also be mentioned over the course of an essay. For example, *"favorable"* and *"well-disposed"* are synonyms for the word *"friendly"* and likely to be good expansions. We used an in-house tool to obtain synonyms from WordNet for each of the prompt words. The lookup involves a word sense disambiguation step to choose the most relevant sense for polysemous words. All the synonyms for the chosen sense of the prompt word are added as expansions.

Distributionally similar words: We also consider as expansions words that appear in similar contexts as the prompt words. For example, *"cordial"*, *"polite"*, *"cheerful"*, *"hostile"*, *"calm"*, *"lively"* and *"affable"* often appear in the same contexts as the word *"friendly"*. Such related words form part of a concept like *'behavioral characteristics of people'* and are likely to appear in a discussion of any one aspect. These expansions could comprise antonyms and other related words too. This idea of word similarity was implemented in work by Lin (1998). Similarity between two words is estimated by examining the degree of overlap of their contexts in a large corpus. We access Lin's similarity estimates using a tool from Leacock and Chodorow (2003) that returns words with similarity values above a cutoff.

Word association norms: Word associations have been of great interest in psycholinguistic research. Participants are given a *target* word and asked to mention words that readily come to mind. The most frequent among these are recorded as *free associations* for that target. They form another interesting category of expansions for our purpose because they are known to be frequently recalled by human subjects for a particular stimulus word. We added the associations for prompt words from a collection of 5000 target words with their associations produced by about 6000 participants (Nelson et al., 1998). Sample associations for the word *"friendly"* include *"smile"*, *"amiable"*, *"greet"* and *"mean"*.

3.2 Weighting of prompt words and expansions

After expansion, the prompt lengths vary between 87 (word associations) and 229 (distributionally similar words) content words, considerably higher than the original average length of 9 and 13 content words. We use a simple weighting scheme[3] to mitigate the influence of noisy expansions. We assign a weight of 20 to original prompt words and 1 to all the expansions. While computing similarity, we use these weight values as the assumed frequency of the word in the prompt. In this case, the term frequency of original words is set as 20 and all expansion terms are considered to appear once in the new prompt.

4 Spelling correction of essay text

Essays written by learners of a language are prone to spelling errors. When such errors occur in the use of the prompt words, prompt-based techniques will fail to identify the essay as on-topic even if it actually is. The usefulness of expansion could also be limited if there are several spelling errors in the essay text. Hence we explored the correction of spelling errors in the essay before off-topic detection.

We use a tool from Leacock and Chodorow (2003) to perform *directed* spelling correction, ie., focusing on correcting the spellings of words most likely to match a given *target* list. We use the prompt words as the targets. We also explore the simultaneous use of spelling correction and expansion. We first obtain expansion words from one of our unsupervised methods. We then use these along with the prompt words for spelling correction followed by matching of the expanded prompt and essay text.

5 Results and discussion

We used our proposed methods on the two essay collections with very short prompts, *Type 3* written by

[3]Without any weighting there was an increase in error rates during development tests. We also experimented with a graph-based approach to term weighting which gave similar results.

advanced test takers and *Type 4* written by learners of English. Table 2 compares the suggested enhancements with the previously proposed method by Higgins et al. (2006). As discussed in Section 2.3, using only the original prompt words, error rates are around 10% for both essay types. For advanced test takers, the false positive rates are lower, around 3%.

Usefulness of expanded prompts All the expansion methods lower the false positive error rates on essays written by learners with almost no increase in the rate of false negatives. On average, the false positive errors are reduced by about 3%. Inflected forms constitute the best individual expansion category. The overall best performance on this type of essays is obtained by combining inflected forms with word associations.

In contrast, for essays written by advanced test takers, inflected forms is the worst expansion category. Here word associations give the best results reducing both false positive and false negative errors; the reduction in false positives is almost 50%. These results suggest that advanced users of English use more diverse vocabulary in their essays which are best matched by word associations.

Effect of spelling correction For essays written by learners, spell-correcting the essay text before comparison (*Spell*) leads to huge reductions in error rates. Using only the original prompt, the false positive rate is 4% lower with spelling correction than without. Note that this result is even better than the best expansion technique–inflected forms. However, for essays written by advanced users, spelling correction does not provide any benefits. This result is expected since these test-takers are less likely to produce many spelling errors.

Combination of methods The benefits of the two methods appear to be population dependent. For learners of English, a spelling correction module is necessary while for advanced users, the benefits are minimal. On the other hand, prompt expansion works extremely well for essays written by advanced users. The expansions are also useful for essays written by learners but the benefits are lower compared to spelling correction. However, for both essay types, the combination of spelling correction and best prompt expansion method (*Spell + best expn.*) is better compared to either of them individually.

	Learners		Advanced	
Method	**FP**	**FN**	**FP**	**FN**
Prompt only	9.73	11.07	2.94	9.06
Synonyms	7.03	12.01	1.39	9.76
Dist.	6.45	11.77	1.63	8.98
WAN	6.33	11.97	**1.59**	**8.74**
Infl. forms	6.25	11.65	2.53	9.06
Infl. forms + WAN	**6.04**	**11.48**	-	-
Spell	5.43	12.71	2.53	9.27
Spell + best expn.	***4.66***	***11.97***	***1.47***	***9.02***

Table 2: Average error rates after prompt expansion and spelling correction

Therefore the best policy would be to use both enhancements together for prompt-based methods.

6 Conclusion

We have described methods for improving the accuracy of off-topic essay detection for short prompts. We showed that it is possible to predict words that are likely to be used in an essay based on words that appear in its prompt. By adding such words to the prompt automatically, we built a better representation of prompt content to compare with the essay text. The best combination included inflected forms and word associations, reducing the false positives by almost 4%. We also showed that spelling correction is a very useful preprocessing step before off-topic essay detection.

References

R.S.J.d. Baker, A.M.J.B. de Carvalho, J. Raspat, V. Aleven, A.T. Corbett, and K.R. Koedinger. 2009. Educational software features that encourage and discourage "gaming the system". In *Proceedings of the International Conference on Artificial Intelligence in Education*.

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.

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

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774.

D. L Nelson, C. L. McEvoy, and T. A. Schreiber. 1998. The University of South Florida word association, rhyme, and word fragment norms, http://www.usf.edu/FreeAssociation/.