Proceedings of HLT-NAACL 2003
Short Papers , pp. 88-90
Edmonton, May-June 2003

A Hybrid Approach to Content Analysis for Automatic Essay Grading

Carolyn P. Rosé, Antonio Roque, Dumisizwe Bhembe, and Kurt VanLehn
LRDC, University of Pittsburgh, 3939 O'hara St., Pittsburgh, PA 15260
rosecp@pitt.edu

Abstract

We present CarmelTC, a novel hybrid text classification approach for automatic essay grading. Our evaluation demonstrates that the hybrid CarmelTC approach outperforms two "bag of words" approaches, namely LSA and a Naive Bayes, as well as a purely symbolic approach.

1 Introduction

In this paper we describe CarmelTC[1], a novel automatic essay grading approach using a hybrid text classification technique for analyzing essay answers to qualitative physics questions inside the Why2 tutorial dialogue system (VanLehn et al., 2002). In contrast to many previous approaches to automated essay grading (Burstein et al., 1998; Foltz et al., 1998; Larkey, 1998), our goal is not to assign a letter grade to student essays. Instead, our purpose is to tally which set of "correct answer aspects" are present in student essays. Previously, tutorial dialogue systems such as AUTO-TUTOR (Wiemer-Hastings et al., 1998) and Research Methods Tutor (Malatesta et al., 2002) have used LSA (Landauer et al., 1998) to perform the same type of content analysis for student essays that we do in Why2. While Bag of Words approaches such as LSA have performed successfully on the content analysis task in domains such as Computer Literacy (Wiemer-Hastings et al., 1998), they have been demonstrated to perform poorly in causal domains such as research methods (Malatesta et al., 2002) because they base their predictions only on the words included in a text and not on the functional relationships between them. Thus, we propose CarmelTC as an alternative. CarmelTC is a rule learning text classification approach that bases its predictions both on features extracted from CARMEL's deep syntactic functional analyses of texts (Rosé, 2000) and a "bag of words" classification of that text obtained from Rainbow Naive Bayes (McCallum and Nigam, 1998). We evaluate CarmelTC in the physics domain, which is a highly causal domain like research methods. In our evaluation we demonstrate that CarmelTC outperforms both Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Rainbow Naive Bayes (McCallum and Nigam, 1998), as well as a purely symbolic approach similar to (Furnkranz et al., 1998). Thus, our evaluation demonstrates the advantage of combining predictions from symbolic and "bag of words" approaches for content analysis aspects of automatic essay grading.

[1]This research was supported by the ONR, Cognitive Science Division under grant number N00014-0-1-0600 and by NSF grant number 9720359 to CIRCLE.

2 Student Essay Analysis

We cast the Student Essay Analysis problem as a text classification problem where we classify each sentence in the student's essay as an expression one of a set of "correct answer aspects", or "nothing" in the case where no "correct answer aspect" was expressed. Essays are first segmented into individual sentence units. Next, each segment is classified as corresponding to one of the set of key points or "nothing" if it does not include any key point. We then take an inventory of the classifications other than "nothing" that were assigned to at least one segment. We performed our evaluation over essays collected from students interacting with our tutoring system in response to the question "Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.", which we refer to as the Pumpkin Problem. Thus, there are a total of six alternative classifications for each segment:

Class 1 After the release the only force acting on the pumpkin is the downward force of gravity.

Class 2 The pumpkin continues to have a constant horizontal velocity after it is released.

Class 3 The horizontal velocity of the pumpkin continues to be equal to the horizontal velocity of the man.

Class 4 The pumpkin and runner cover the same distance over the same time.

Class 5 The pumpkin will land on the runner.

Class 6 Sentence does not adequately express any of the above specified key points.

Often what distinguishes sentences from one class and another is subtle. For example, "The pumpkin's horizontal velocity, which is equal to that of the man when he released it, will remain constant." belongs to Class 2. However, it could easily be mistaken for Class 3 based on the set of words included, although it does not express that idea since it does not address the relationship between the pumpkin's and man's velocity *after* the release. Similarly, "So long as no other horizontal force acts upon the pumpkin while it is in the air, this velocity will stay the same.", belongs to Class 2 although looks similar on the surface to either Class 1 or 3. Nevertheless, it does not express the required propositional content for either of those classes. The most frequent problem is that sentences that express most but not all of the content associated with a required point should be classified as "nothing" although they have a lot of words in common with sentences from the class that they are most similar to. Similarly, sentences like "It will land on the ground where the runner threw it up." contain all of the words required to correctly express the idea corresponding to Class 5, although it does not express that idea, and in fact expresses a wrong idea. These very subtle distinctions pose problems for "bag of words" approaches since they base their decisions only on which words are present regardless of their order or the functional relationships between them.

The hybrid CarmelTC approach induces decision trees using features from a deep syntactic functional analysis of an input text as well as a prediction from the Rainbow Naive Bayes text classifier (McCallum and Nigam, 1998). Additionally, it uses features that indicate the presence or absence of words found in the training examples. From these features CarmelTC builds a vector representation for each sentence. It then uses the ID3 decision tree learning algorithm (Quinlin, 1993) to induce rules for identifying sentence classes based on these feature vectors.

From CARMEL's deep syntactic analysis of a sentence, we extract individual features that encode functional relationships between syntactic heads (e.g., (subj-throw man)), tense information (e.g., (tense-throw past)), and information about passivization and negation (e.g., (negation-throw +) or (passive-throw -)). Syntactic feature structures produced by the grammar factor out those aspects of syntax that modify the surface realization of a sentence but do not change its deep functional analysis, including syntactic transformations such as passivization and extraction. These deep functional relationships give CarmelTC the information lacking on Bag of Words approaches that is needed for effective content analysis in highly causal domains, such as research methods or physics.

3 Evaluation

We conducted an evaluation to compare the effectiveness of CarmelTC at analyzing student essays in comparison to LSA, Rainbow, and a purely symbolic approach similar to (Furnkranz et al., 1998), which we refer to here as CarmelTCsymb. CarmelTCsymb is identical to CarmelTC except that it does not include in its feature set the prediction from Rainbow. We conducted our evaluation over a corpus of 126 previously unseen student essays in response to the Pumpkin Problem described above, with a total of 500 text segments, and just under 6000 words altogether. Each text segment was hand tagged by at least two coders, and conflicts were resolved at a consensus meeting. Pairwise Kappas between our three coders computed over initial codings of our data was always above .75.

The LSA space used for this evaluation was trained over three first year physics text books. The Rainbow models used to generate the Rainbow predictions that are part of the feature set provided to CarmelTC were trained over a development corpus of 248 hand tagged example sentences extracted from a corpus of human-human tutoring dialogues, just like those included in the 126 essays mentioned above. However, when we evaluated the performance of Rainbow for comparison with CarmelTC, LSA, and the symbolic approach, we ran a 50 fold cross validation evaluation using the complete set of examples in both sets (i.e., the 248 sentences used to train the Rainbow models used to by CarmelTC as well as the 126 essays) so that Rainbow would have access to the exact same training data as CarmelTC, to make it a fair comparison between alternative machine learning approaches. On each iteration, we randomly selected a subset of essays such that the number of text segments included in the test set were greater than 10 but less than 15 and then training Rainbow using the remaining text segments. Thus, CarmelTC uses the same set of training data, but unlike the other approaches, it uses its training data in two separate parts, namely one to train the Rainbow models it uses to produce the Rainbow prediction that is part of the vector representation it builds for each text and one to train the decision trees. This is because for CarmelTC, the data for training Rainbow must be separate from that used to train the decision trees so the decision trees are trained from a realistic distribution of assigned Rainbow classes based on its performance on unseen data rather than on

Figure 1: **This Table compares the performance of the 3 alternative approaches**

Approach	Precision	Recall	False Alarm Rate	F-Score
LSA	93%	54%	3%	.70
Rainbow	81%	73%	9%	.77
CarmelTCsymb	88%	72%	7%	.79
CarmelTC	90%	80%	8%	.85

Rainbow's training data. Thus, for CarmelTC, we also performed a 50 fold cross validation, but this time only over the set of 126 example essays not used to train the Rainbow models used by CarmelTC.

Note that LSA works by using its trained LSA space to construct a vector representation for any text based on the set of words included therein. It can thus be used for text classification by comparing the vector obtained for a set of exemplar texts for each class with that obtained from the text to be classified. We tested LSA using as exemplars the same set of examples used as Rainbow training data, but it always performed better when using a small set of hand picked exemplars. Thus, we present results here using only those hand picked exemplars. For every approach except LSA, we first segmented the essays at sentence boundaries and classified each sentence separately. However, for LSA, rather than classify each segment separately, we compared the LSA vector for the entire essay to the exemplars for each class (other than "nothing"), since LSA's performance is better with longer texts. We verified that LSA also performed better specifically on our task under these circumstances. Thus, we compared each essay to each exemplar, and we counted LSA as identifying the corresponding "correct answer aspect" if the cosine value obtained by comparing the two vectors was above a threshold. We used a threshold value of .53, which we determined experimentally to achieve the optimal f-score result, using a beta value of 1 in order to treat precision and recall as equally important.

Figure 1 demonstrates that CarmelTC out performs the other approaches, achieving the highest f-score, which combines the precision and recall scores into a single measure. Thus, it performs better at this task than two commonly used purely "bag of words" approaches as well as to an otherwise equivalent purely symbolic approach.

References

J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M. D. Harris. 1998. Automated scoring using a hybrid feature identification technique. In *Proceedings of COLING-ACL'98*, pages 206–210.

P. W. Foltz, W. Kintsch, and T. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307.

J. Furnkranz, T. Mitchell Mitchell, and E. Riloff. 1998. A case study in using linguistic phrases for text categorization on the www. In *Proceedings from the AAAI/ICML Workshop on Learning for Text Categorization*.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

L. Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of SIGIR*.

K. Malatesta, P. Wiemer-Hastings, and J. Robertson. 2002. Beyond the short answer question with research methods tutor. In *Proceedings of the Intelligent Tutoring Systems Conference*.

A. McCallum and K. Nigam. 1998. A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Classification*.

J. R. Quinlin. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers: San Mateo, CA.

C. P. Rosé. 2000. A framework for robust semantic interpretation. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 311–318.

K. VanLehn, P. Jordan, C. P. Rosé, and The Natural Language Tutoring Group. 2002. The architecture of why2-atlas: a coach for qualitative physics essay writing. Proceedings of the Intelligent Tutoring Systems Conference.

P. Wiemer-Hastings, A. Graesser, D. Harter, and the Tutoring Res earch Group. 1998. The foundations and architecture of autotutor. In B. Goettl, H. Halff, C. Redfield, and V. Shute, editors, *Intelligent Tutoring Systems: 4th International Conference (ITS '98)*, pages 334–343. Springer Verlag.