NLTP and ML technique for Text categorization

[1]Jayababu Yasarapu , [2]Rajesh M V , [3]B S N V Ramana Murthy, [4]Soma Sekhar Turangi

Department of CSE, PRAGATI Engineering College(Autonomous), Surampalem, A.P, India.

ABSTRACT

Text classification is the process of classifying documents into predefined categories based on their content. Existing supervised learning algorithms to automatically classify text need sufficient documents to learn accurately. This paper presents a new algorithm for text classification that requires fewer documents for training. Instead of using words, word relation i.e association rules from these words is used to derive feature set from preclassified text documents. The concept of Naïve Bayes classifier is then used on derived features and finally only a single concept of Genetic Algorithm has been added for final classification. Experimental results show that the classifier build this way is more accurate than the existing text classification systems.

INTRODUCTION

In this paper we describe CarmelTC, a novel hybrid text classification approach for analyzing essay answers to qualitative physics questions. In our evaluation we demonstrate that the novel hybrid CarmelTC approach outperforms both Latent Semantic Analysis (LSA) (Landauer et al., 1998; Laham, 1997) and Rainbow (McCallum, 1996; McCallum and Nigam, 1998), which is a Naive Bayes approach, as well as a purely symbolic approach similar to (Furnkranz et al., 1998). Whereas LSA and Rainbow are pure “bag of words” approaches,CarmelTC is a rule learning approach where rules for classifying units of textrely on features extracted from a syntactic analysis of that text as well as on a “bag of words” classification of that text. Thus, our evaluation demonstrates the advantage of combining predictions from symbolic and “bag of words” approaches for text classification. Similar to (Furnkranz et al., 1998), neither CarmelTC nor the purely symbolic approach require any domain specific knowledge engineering or text annotation beyond providing a training corpus of texts matched with appropriate classifications, which is also necessary for Rainbow, and to a much lesser extent for LSA. CarmelTC was developed for use inside of the Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2002; Graesser et al., 2002) for the purpose of grading short essays written in response to questions such as “Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.” This is an appropriate task domain for pursuing questions about the benefits of tutorial dialogue for learning because questions like this one are known to elicit robust, persistent misconceptions from students, such as “heavier objects exert more force.” (Hake, 1998; Halloun and Hestenes, 1985). In Why2-Atlas, a student first types an essay answering a qualitativephysics problem. A computer tutor then engages the student in a natural language dialogue to provide feedback, correct misconceptions, and to elicit more complete explanations. The first version of Why2-Atlas was deployed and evaluated with undergraduate students in the spring of 2002; the system is continuing to be actively developed (Graesser et al., 2002). In contrast to many previous approaches to automated essay grading (Burstein et al., 1998; Foltz et al., 1998; Larkey, 1998), ourgoal 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. For example, we expect satisfactory answers to the example question above to include a detailed explanation of how Newton’s first law applies to this scenario. From Newton’s first law, the student should infer that the pumpkin and the man will continue at the same constant horizontal velocity that they both had before the release. Thus, they will always have the same displacement from the point of release. Therefore, after the pumpkin rises and falls, it will land back in the man’s hands. Our goal is to coach students through the process of constructing good physics explanations. Thus, our focus is on the physics content and not the quality of the student’s writing, in contrast to (Burstein et al., 2001).

LITERATURE SERVUY

In this paper we describe CarmelTC, a novel hybrid text classification approach for analyzing essay answers to qualitative physics questions. In our evaluation we demonstrate that the novel hybrid CarmelTC approach outperforms both Latent Semantic Analysis (LSA) (Landauer et al., 1998; Laham, 1997) and Rainbow (McCallum, 1996; McCallum and Nigam, 1998), which is a Naive Bayes approach, as well as a purely symbolic approach similar to (Furnkranz et al., 1998). Whereas LSA and Rainbow are pure “bag of words” approaches,CarmelTC is a rule learning approach where rules for classifying units of textrely on features extracted from a syntactic analysis of that text as well as on a “bag ofwords” classification of that text. Thus, our evaluation demonstrates the advantage of combining predictions from symbolic and “bag of words” approaches for text classification. Similar to (Furnkranz et al., 1998), neither CarmelTC nor the purely symbolic approach require any domain specific knowledge engineering or text annotation beyond providing a training corpus of texts matched with appropriate classifications, which is also necessary for Rainbow, and to a much lesser extent for LSA. CarmelTC was developed for use inside of the Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2002; Graesser et al., 2002) for the purpose of grading short essays written in response to questions such as “Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.” This is an appropriate task domain for pursuing questions about the benefits of tutorial dialogue for learning because questions like this one are known to elicit robust, persistent misconceptions from students, such as “heavier objects exert more force.” (Hake, 1998; Halloun and Hestenes, 1985). In Why2-Atlas, a student first types an essay answering a qualitativephysics problem. A computer tutor then engages the student in a natural language dialogue to provide feedback, correct misconceptions, and to elicit more complete explanations. The first version of Why2-Atlas was deployed and evaluated with undergraduate students in the spring of 2002; the system is continuing to be actively developed (Graesser 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. For example, we expect satisfactory answers to the example question above to include a detailed explanation of how Newton’s first law applies to this scenario. From Newton’s first law, the student should infer that the pumpkin and the man will continue at the same constant horizontal velocity that they both had before the release. Thus, they will always have the same displacement from the point of release. Therefore, after the pumpkin rises and falls, it will land back in the man's hands. Our goal is to coach students through the process of constructing good physics explanations. Thus, our focus is onthe physics content and not the quality of the student's writing, in contrast to (Burstein et al., 2001). 2 Student EssayAnalysis 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 "correctanswer aspect" was expressed. After a student attempts an initial answer to the question, the system analyzes the student's essay to assess which key points are missing from the student's argument. The system then uses its analysis of the student's essay to determine which help to offer that student. In order to do an effective job at selecting appropriate interventions for helping students improve their explanations, the system mustperform a highly accurate analysis of the student's essay. Identifying key points as present in essays when they are not (i.e., false alarms), cause the system to miss opportunities to help students improve their essays. On the other hand, failing to identify key points that are indeed present in student essays causes the system to offer help where it is not needed, which can frustrate and even confuse students. A highly accurate inventory of the content of student essays is required in order to avoid missing opportunities to offer needed instruction and to avoid offering inappropriate feedback, especially as the completeness of student essays increases (Rose´ et al., 2002a; Rose´ et al., 2002c). In order to compute which set of key points, i.e., "correct answer aspects", are included in a student essay, we first segment the essay at sentence boundaries. Note that run-on sentences are broken up. Once an essay is segmented, each segment is classified as corresponding to one of the set of key points or "nothing" if it does notinclude any key point. We then take an inventory of the classifications other than "nothing" that were assigned to at leastone segment. Thus, our approach is similar in spirit to that taken in the AUTO- TUTOR system (WiemerHastings et al., 1998), where Latent Semantic Analysis (LSA) (Landauer et al., 1998; Laham, 1997) was used to tally which subset of "correct answer aspects" students included in their natural language responses to shortessay questions about computer literacy. 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 totalof six alternative classifications for each segment: Class 1 Sentence expresses the idea that after the release the only force acting on the pumpkin is the downward force of gravity. Class 2 Sentence expresses the idea that the pumpkin continuesto have a constant horizontal velocity after it is released. Class 3 Sentence expresses the idea that the horizontal velocity of the pumpkin continues to be equal to the horizontal velocity of the man. Class 4 Sentence expresses the idea that the pumpkin and runner cover the same distance over the same time. Class 5 Sentence expresses the idea that the pumpkin will land on the runner. Class 6 Sentence does not adequately express any of the above specified key points. Note that this classification task is strikingly different from those typically used for evaluating text classification systems. First, these classifications represent specific whole propositions rather than general topics, such as those used for classifying web pages (Craven et al., 1998), namely "student", "faculty", "staff", etc. Secondly,the texts are much shorter, i.e., one sentence in comparison with a whole web page, which is a disadvantage for “bag of words” approaches. In some cases what distinguishes sentences from one class and sentences from another class is very subtle. For example, “Thus, the pumpkin’s horizontal velocity, which is equal to that of the man when he released it, will remain constant.” belongs to Class 2 although it could easily be mistaken for Class 3. 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. A related problem is that sentences that should be classified as “nothing” may look very similar on the surface to sentences belonging to one or more of the other classes. For example, “It will land on the ground where the runner threw it up.” contains all of the words required to correctly express the idea corresponding to Class 5, although it does not express this idea, and in fact expresses a wrong idea. These very subtle distinctions also pose problems for “bag of words” approaches since they base their decisions only on which words are present regardless of theirorder or the functional relationships between them. That might suggest that a symbolic approach involving syntactic and semantic interpretation might be more successful. However, while symbolic approaches can be more precise than “bag of words” approaches, they are also more brittle. And approaches that rely both on syntactic and semantic interpretation require a larger knowledge engineering effort as well.

RELATED WORK

CarmelTC is most similar to the text classification approach described in (Furnkranz et al., 1998). In the approach described in (Furnkranz et al., 1998), features that note the presence or absence of a word from a text as well as extraction patterns from AUTOSLOG-TS (Riloff, 1996) form the feature set that are input to the RIPPER (Cohen, 1995), which learns rules for classifying texts based on these features. CarmelTC is similar in spirit in terms of both the sorts of features used as well as the general sort of learning approach. However, CarmelTC is different from (Furnkranz et al., 1998) in several respects. Where (Furnkranz et al., 1998) make use of AUTOSLOG-TS extraction patterns, CarmelTC makes use of features extracted from a deep syntactic analysis of the text. Since AUTOSLOG-TS performs a surface syntactic analysis, it would assign a different representation to all aspects of these texts where there is variation in the surface syntax. Thus, the syntactic features extracted from our syntactic analyses are more general. For example, for the sentence “The force was applied by the man to the object”, our grammar assigns the same functional roles as for “The man applied the force to the object” and also for the noun phrase “the man that applied the force to the object”. This would not be the case for AUTOSLOGTS. Like (Furnkranz et al., 1998), we also extract word features that indicate the presence or absence of a root form of a word from the text. However, in contrast for CarmelTC one of the features for each training text that is made available to the rule learning algorithm is the classification obtained using the Rainbow Naive Bayes classifier (McCallum, 1996; McCallum and Nigam, 1998). Because the texts classified with CarmelTC are so much shorter than those of (Furnkranz et al., 1998), the feature set provided to the learning algorithm was small enough that it was not necessary to use a learning algorithm as sophisticated as RIPPER (Cohen, 1995). Thus, we used ID3 (Mitchell, 1997; Quinlin, 1993) instead with excellent results. Note that in contrastto CarmelTC, the (Furnkranz et al., 1998) approach is purely symbolic. Thus, all of its features are either word level features or surface syntactic features. Recent work has demonstrated that combining multiple predictors yields combined predictors that are superior to the individual predictors in cases where the individual predictors have complementary strengths and weaknesses (Larkey and Croft, 1996; Larkey and Croft, 1995). We have argued that this is the case with symbolic and “bag of words” approaches. Thus, we have reason to expect a hybrid approach that makes a prediction based on a combination of these single approaches would yield better results than either of these approaches alone. Our results presented in Section 5 demonstrate that this is true. Other recent work has demonstrated that symbolic and “Bag of Words” approaches can be productively combined. For example, syntactic information can be used to modify the LSA space of a verb in order tomake LSA sensitive to different word senses (Kintsch, 2002). However, this approach has only been applied to the analysis of mono-transitive verbs. Furthermore, it has never been demonstrated to improve LSA’s effectiveness at classifying texts. In the alternative Structured Latent Semantic Analysis (SLSA) approach, hand-coded subject-predicate information was used to improve the results obtained by LSA for text classification (Wiemer-Hastings and Zipitria, 2001), but no fully automated evaluation of this approach has been published. In contrast to these two approaches, CarmelTC is both fully automatic, in that the symbolic features it uses are obtained without any hand coding whatsoever, and fully general, in that itapplies to the full range of verb subcategorization frames covered by the COMLEX lexicon, not only mono-transitive verbs. In Section 5 we demonstrate that CarmelTC outperforms both LSA and Rainbow, two alternative bag of words approaches, on the task of student essay analysis.

Proposed Algorithm

```
n = number of class, m = number of associated sets
1.For each class i = 1 to n do
2.Set pval = 0, nval = 0, p = 0, n = 03.For each set s = 1 to m do
4.If the probability of the class (i) for theset (s) is maximum then increment pval elseincrement nval
5.If 50% of the associated set s is matchedwith the keywords set do step 6 else do step7
6.If maximum probability matches theclass i then increment p
7.If maximum probability does not matchthe class i increment n
8.If (s<=m) go to step 3
9.Calculate the percentage of matching inpositive sets for the class i
10.Calculate the percentage of notmatching in negative sets for the class i
11.Calculate the total probability as the summation of the results obtained from step 9 and 10and also the prior probability of the class i inset s
12.If (i<=n) go to step 1
13.Set the class having the maximum
```

probability
value as the result

CONCLUSION

This paper presented a new hybrid technique for text classification. The existing algorithms require more data for training as well as the computational time of these algorithms also increases. Incontrast to the existing algorithms, the proposed hybrid algorithm requires less training data and less computational time. In spite of the randomly chosen training set we achieved 78% accuracy for 50% training data. Though 85% accuracy was observed in 30% training data, a class could not be classified, so we dropped this position and increased training data set for more acceptable result.

REFERENCES

[1] Agarwal, R., Mannila H., Srikant R., Toivonan H., Verkamo A., "A Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining, 1996. [2] Anwar M. Hossain, Mamunur M. Rashid, Chowdhury Mofizur Rahman, "A New Genetic Algorithm Based Text Classifier," In Proceedings of International Conference on Computer and Information Technology, NSU, pp. 135-139, 2001. [3] Canasai Kruengkrai, Chuleerat Jaruskulchai, "A Parallel Learning Algorithm for Text Classification," The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2002), Canada, July 2002. [4] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, Kamruzzaman S. M, "Text Classification Using the Concept of Association Rule of Data Mining," In Proceedings of International Conference on Information Technology, Kathmandu, Nepal, pp 234- 241, May 23-26, 2003. [5] Eshita Sharmin, Ayesha Akhter, Chowdhury Mofizur Rahman, "Genetic Algorithm for Text Categorization," In Proceedings of International Conference on Computer and Information Technology, BUET, pp. 80-85, December, 1998. [6] Jinyan Li, Thomas Manoukian, Guozhu Dong, and Kotagiri Ramamohanarao. Incremental Maintenance of the Border of the Space of Emerging Patterns. Data Mining and Knowledge Discovery, 9 (1): 89- 116, July 2004 [7] Lewis, D., and Ringuette, M., "A Comparison of Two Learning Algorithms for Text Categorization," In Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, 1994.