IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5013

Automated Essay Grading via Text Classification

Salvatore Valenti
DIIGA - Università Politecnica delle Marche - Ancona - Italy, valenti@diiga.univpm.it

Alessandro Cucchiarelli
DIIGA - Università Politecnica delle Marche - Ancona - Italy, Alex@diiga.univpm.it

INTRODUCTION

Essays are considered by many researchers as the most useful tool to assess learning outcomes implying a) the ability to recall, organize and integrate ideas, b) the ability both to express oneself in writing and c) to supply more than identify interpretation and application of data.

One of the difficulties of grading essays is represented by the perceived subjectivity of the grading process. Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness. This issue may be faced through the adoption of tools for Automated Essay Grading (AEG). An AEG system would at least be consistent in the way it scores essays, and enormous cost and time savings could be achieved if the system can be shown to grade essays within the range of those awarded by human assessors. Moreover, an AEG system would be an extremely useful and valuable tool for distance learning students needing to practice self assessment on those topics that could not be easily covered via closed answer tests.

Page in (1996) introduced a distinction between grading essays for content and for style, where the former refers loosely to what an essay says, while the latter to "syntax and mechanics and diction and other aspects of the way it is said". In the current literature on AEG systems papers reporting experiments with systems aimed to evaluate essays primarily for content or for style, are discussed. Furthermore, systems aimed to evaluate essays taking in account both aspects are reported too (Valenti et al., 2003).

Three different criteria have been discussed to measure the performance of AEG systems: accuracy of the results, multiple regression correlation and percentage of agreement between grades produced by the systems those assigned by human experts (Valenti et al., 2003).

This paper is aimed to discuss the design of an AEG system that we are developing at the Università Politecnica delle Marche. The system will be initially devoted to grade essays for content, and will be based on text classification techniques defined in the context of our research in Natural Language Processing (Cucchiarelli 2001, Velardi 2000).

Text Classification (TC) is the problem of assigning predefined categories to free text documents. The approach adopted relays on the availability of a large collection of documents that is used to train the classification system and to build the classes profiles. In our approach, the TC system will be trained on a collection of human-graded essays to create models of grading classes. Then, the obtained model will be used to classify previously unseen essays. The performances of the AEG system will be measured by comparing the percentage of agreement between the produced grades and those assigned by human experts to the unseen essays.

Since no public domain collection of essays is actually available, this paper reports on our solution to solve this problem, too.

The paper is organized as follows: in the first section some background information on text classification is provided. Then, our approach to automated essay marking via text classification along with the outline of the system under development, will be provided.

FUNDAMENTALS OF TEXT CLASSIFICATION

Text Classification (TC) is the problem of assigning predefined categories to free text documents. Typically, the classifiers adopt two-phased machine learning algorithms: a training phase, in which the grading rules are acquired using various algorithms, and a testing phase, in which the rules gathered in the first step are used to determine the most probable grade for a particular essay.

Less informally, TC is the task of assigning documents to a set of predefined categories, and can be modelled as follows. Given a set of conceptual classes $C = (c_1, c_2, \ldots, c_h)$ related to the topics of interest and a set of training documents $D = (d_1, d_2, \ldots, d_k)$ each labelled according to the classes it belongs to, build a decision function f able to assign the correct classes to each document, i.e. f : D ®ð 2^c.

The function f can be further applied to newly incoming documents, to classify them in one or more classes of C.

According to the current research on this field, the design of a TC system consists of a set of subtasks, namely features design, features weighting, similarity estimation, inference and testing (Moschitti, 2003), that will be further discussed in this section.

Features Design

The training documents are usually represented as feature vectors, i.e. t-uples of values $X=(x_1, x_2, \ldots, x_n)$, being x_j the numeric value that feature j takes on for document X. For example, if feature j is a word, x_j could be the value related to the frequency of j in X. The selection of the n relevant features to be used in vectors definition has a strong influence on the classifier performances and is a very critical task. As words in the document are usually considered as basic unit of information, a sub-set of them (ignoring uninformative terms like articles, pronouns, adverbs known as stop words) are candidate as features. The set of candidate words are then normalized, through stemming (removing common suffixes from words) or lemmatisation (reducing each word to its base form).

Feature Weighting

Features could be more or less representative in documents. Roughly speaking, the higher in a document the frequency of a word is, the more it characterize the document itself; on the contrary, the wider the occurrence of a word in the entire set of training documents, the less its relevance for each single document. Many different schemes have been devised for the estimation of the weight x_j as for instance *: IDF ´ð TF* (Salton 1991), *log(TF) ´ IDF* (Inner et al. 1995) and *IWF* (Basili et al. 1999) and different systems may benefit from the use of different weighting schemes.

Once the appropriate policy has been chosen, the weights for the class profile can be obtained. The class profile is a vector $c_i = (w_1^i, w_2^i, \ldots, w_n^i)$, and the value of each element w_j^i is the result of the evaluation of the relevance of feature j in the training set documents. Different policies can be applied to compute each w_j^i: from the simplest one

$$w_j^i = \sum_{h \in T_i} w_j^h$$

that sums up the weights of feature j in all the training documents

T_i belonging to class c_i, to the more complex ones, that also use negative evidence provided by documents not belonging to the class (see, for example, Rocchio's algorithm (Rocchio, 1971)).

Similarity Estimation

Having both the classes' profiles and the feature vector of a previously unseen document represented in the same manner, it is then possible to estimate the similarity among the document and each class in *C*. The estimation is usually made by using operations in the space of features. The most popular operator is the cosine of the angle between the vector c_i and the vector $d = (w_1^d, w_2^d, \ldots, w_n^d)$ representative of the document, applied to estimate the similarity s_{id} as follows:

$$s_{id} = \cos(c_i, d) = \sum_{j=1}^{n} w_j^i w_j^d$$

The above formula, computed for each class profile c_i, is the basis for the further classification of *d* over the classes in *C*.

Inference

A decision function is usually applied over the similarity scores s_{id} to assign an incoming document to one or several target classes. This is carried out by defining a threshold sð so that only the documents having s_{id} ³ð sð are classified in c_i. Different strategies can be used to define sð (Yang, 1999). A value of threshold $sð_i$ can be assigned to each class in *C*, as a probabilistic measure related to the risk of a document misclassification in a class (Scut), or to the probability $prob(c_i \mid T)$ of documents classified in c_i in the training set *T* (Rcut). Moreover, the value *k* of the average number of classes valid for a generic document *d* can be used as a fixed threshold, estimated usually over the training set (Rcut); in this case, the first *k* ranked classes having positive value of s_{id} are assigned to document *d*.

Testing

The accuracy of the classification process defined by the previous steps is then evaluated over a *test set* of pre-labelled documents, disjoint by that used for training (*training set*). The correctness of classification is estimated by comparing the distance between human classification (given by the documents labels in the test set, usually assigned by human experts), and the output of the inference phase. Many different scores can be used for this estimation (Yang, 1999). They range from the classic *recall* and *precision*, (respectively, the ratio between the number of documents correctly classified in c_i by the system and the documents classified in c_i by humans, and the ratio between the number of documents correctly classified in c_i by the system and all the documents it is assigned to c_i), to more complex ones like F_1 (that balances in a single measure *recall* and *precision*) or *Break Even Point* (the measure of classification performance when *recall* and *precision* have the same value). Some of these measures may be misleading when examined alone, so the use of multiple scores is a common practice.

FROM TEXT CLASSIFICATION TO ESSAY GRADING

Essay grading is a task that may be accomplished by considering at least two aspects: the style of the essay and its content. Our research is not concerned with the evaluation of style, even if we strongly believe that a system for automated essay grading must cope with this aspect. Currently, the efforts are concentrating on the definition of a general methodology to grade the content of an essay by using TC techniques.

The basic idea is to consider an essay as a document to be classified in one or more classes, each being an expression of a different grade, in a ranging running from 'excellent' to 'poor'. Given the characteristics of the TC techniques defined in the previous section, by focusing the representation of a single document on the relevant words it contains, this approach promises to be well suited for grading essays in domains characterized by a specific terminology, as expression of the surface appearance of relevant domain concepts. Under these circumstances, some relevant parameters used in grading by human expert, such as *completeness*, *correctness* and *use of proper terminology*, are evaluated by analysing essay terms. Presence or absence of terminological elements, along with their frequencies in text, is the basis for grading each single essay. TC systems use the same metric to grade an essay with respect to a set of grading classes' profiles. Roughly speaking, the c_i profile obtained in the training phase defines the number of domain terms used in documents belonging to the class (through features with $w_j^i \neq 0$) and the relevance of these terms (through the values of w_j^i), so modelling the human definition of grading classes on a terminological perspective.

In order to conduct some experiments a corpus of essays is needed. The term corpus has been used to designate a body of natural language data which can be used as a basis for linguistic research (Leech, 1997). Currently, the term has been applied to a body of language text that exist in electronic format. According to the discussion regarding the principles underlying the TC techniques, our AEG is based on a training phase, in which the grading rules are acquired using various algorithms, and a testing phase, in which the rules gathered in the first step are used to determine the most probable grade for a particular essay. As we started our research, we tried to find out a public domain corpus of essays already marked by human graders. The difficulty of obtaining a test bed has been outlined by other authors (Larkey, 2003; Christie, 2003). This is the reason why we decided to start the construction of an ad-hoc corpus that comes from the essays obtained from the summative assessment of a course in Economics for Business Management that is offered at our University. The essays are written in Italian, and their content has been graded by a human grader considering, as main parameters, the use of proper terminology, the completeness and the correctness, along with other minor aspects. The grades range from A to E. Thus, the corpus is constituted by a collection of essays handed by the students and is annotated with some additional information including a reference to the question asked, the name of the grader, the grade assigned, the date of the test and the topic covered.

The TC model we adopted has been developed in partnership with the Linguistic Computing Group at the Dipartimento di Informatica, Università di Roma "La Sapienza", and is based on the following approach:

Sub-task	Solution adopted
Feature Design	Lemmatisation of essays words. Filtering of predefined class of stop word.
Feature Weighting	Use of $IDF \times TF$ as weight algorithm. Extraction of class profiles through Rocchio's formula.
Similarity Estimation	Use of the cosine operator.
Inference	Application of *Rcut* strategy.
Testing	Evaluation of results obtained over the test set by using *precision*, *recall* and F_1 measures.

FINAL REMARKS

This paper presents the overview of an automated approach to the problem of grading essays for content, via Text Classification. All the modules implementing the different subtasks that need to be performed to automatically classify text documents have been already implemented. Our effort is currently focused in linking together the various modules in order to obtain a system that may be used to support essay grading. At the same time, the construction of the corpus is in progress, and in the near future we should be able to perform some preliminary experiments.

As long as the performances of the system are to some extent comparable to those of the human grader, we have a number of issues to cover, including to a) identify some metrics for the assessment of the style of the essays, b) compare the performances of our system with those claimed by other authors (Valenti, 2003), c) extend the size of the corpus and identify some techniques for keeping in account multiple grades assigned by different human graders to the same essay.

ACKNOWLEDGEMENTS

We would like to thank Prof. Donato Iacobucci of DIIGA - Univ. Politecnica delle Marche - Ancona, for providing us the first batch of

essays on Economics for Business Management that will compose the corpus we are building.

REFERENCES

R. Basili, A. Moschitti, M.T. Pazienza. (1999). “A text classifier based on linguistic processing”, in Proceedings of the 9th International Joint Conference of Artificial Intelligence (IJCAI 1999), Machine Learning for Information Filtering, Stockholm (Swedeen).

Christie J. R. (2003). Email communication with author. 14th April.

Cucchiarelli A.,Velardi P. (2001). “Unsupervised named Enity Recognition Using Syntactic and Semantic Contextual Evidence”, Computational Linguistics, 27 (1), 123-131.

Inner D. J., Lewis D. D. and Ahn D. D. (1995). “Text categorization of low quality images”, in Proceedings of SDAIR-95, pp. 301-315, Las Vegas, USA.

Larkey L. S. (1998). “Automatic essay grading using text categorization techniques” , in Proceedings of the 21st ACM/SIGIR (SIGIR-98), 90-96. ACM.

Larkey L. S. (2003). “Email communication with author”, 15th April.

Moschitti A. (2003). Natural language processing and automated text categorization. PhD thesis, Dept. of Computer Science, Systems and Production, University of Rome “Tor Vergata”.

Page E. B. (1996). “Grading Essay by Computer: Why the Controversy?”, Handout for NCME Invited Symposium.

Page E. B. (1994). “New Computer Grading of Student Prose, Using Modern Concepts and Software” , Journal of Experimental Education , 62(2), 127-142.

Rocchio J. J. (1971). “Relevance feedback in information retrieval”, in G. Salton (ed.), The SMART retrieval system – Experiments in automatic document processing, 313-323. Prentice Hall, Englewood Cliffs, NJ, USA.

Salton G. (1991). “Development in automatic text retrieval”, Science, V. 253, 974-980.

Valenti S., Neri F., Cucchiarelli A. (2003). “An overview of Current Approaches to Automated Essay Grading”, Journal of Information Technology Education, Vol 2, 319-330. http://jite.org

Velardi P., Cucchiarelli A. (2000) “A Theoretical Analysis of Context-based Learning Algorithms for Word Sense Disambiguation”, 14th European Conference on Artificial Intelligence, ECAI-2000, Berlin.

Williams R. (2001). “Automated Essay Grading: an Evaluation of Four Conceptual Models”, in A. Hermann and M.M. Kulski (eds), “Expanding Horizons in Teaching and Learning”, Proc. of the 10th Annual Teaching and Learning Forum, Perth: Curtin University of Technology.

Yang Y. (1999). “An evaluation of statistical approaches to text categorization”, Information Retrieval Journal, Vol.1, N.1/2, 69-90, Kluwer Academic Publishers. The Netherlands.