Identifying Important Factors in Essay Grading Using Machine Learning

Victor D. O. Santos[1]
Iowa State University

Marjolijn Verspoor[2]
University of Groningen

John Nerbonne[3]
University of Groningen

Writing fluently and accurately is a goal for many advanced learners of English. Assigning grades to students' writing is part of the educational process, and some grades, e.g., in standardized tests, even serve as a qualification for entering universities. This chapter aims to identify factors which correlate strongly with essay grades and English proficiency level. Because we would like to contribute to Automated Essay Scoring (AES), we focus on candidate factors that might easily be automated. The study is based on an existing database containing 81 variables (many of which hand-coded) in spontaneously, short written texts by 481 Dutch high school learners of English. All texts were holistically scored on a proficiency level from 1 to 5 by a team of experts. Another question we investigate is which machine learning algorithm (from a subset of those present in the WEKA [4] data mining/machine learning software) provides the best classification accuracy in terms of predicting the English proficiency level of the essays. Logistic Model Tree (LMT), which uses logistic regression, achieves the best accuracy rates (both in terms of precise accuracy and adjacent accuracy) when compared to human judges. The aim of this chapter is thus to build a bridge between applied linguistics and language technology in order to find features (factors) that determine essay grades, with a view to future implementation in an AES system.

Key words: machine learning, English proficiency level, essay scoring, attribute selection, language testing

[1] vdsantos@iastate.edu
[2] m.h.verspoor@rug.nl
[3] j.nerbonne@rug.nl
[4] WEKA stands for Waikato Environment for Knowledge Analysis.

1. Introduction

Writing fluently and accurately is important in numerous jobs and professions, and essay writing is an important exercise in learning to write fluently and accurately. Writing proficiency is important in that it is often tested and graded as part of admissions procedures for higher education and professional education. The TOEFL exam and others such as the Cambridge exams and IELTS are high-stakes tests, which might have a large impact on test takers, such as allowing them or not to enter a graduate program at an American university, depending on their score. Other essay grades might be used for purposes that do not impact test takers' lives so strongly, such as helping teachers decide in which English class at a language center a certain student should be placed.

Identifying the factors that determine essay grades is difficult because there are many potential ones which have been proposed, many of which may in fact not be relevant to the construct at hand, that is, English proficiency. One might decide to focus on just a few factors, but, typical real-world data includes various attributes, only a few of which are actually relevant to the true target concept (Landwehr, Hall, & Frank, 2005). In our case, the target concept, or construct at hand, is English proficiency.

In this chapter, we would like to use real-world data—texts written by L2 learners of English, which are coded for 81 linguistic variables— and identify which set of factors is actually relevant to the true target concept: the proficiency level of the learners. To do so we first investigate to what extent machine learning algorithms and techniques, such as those implemented in the widely used WEKA package (University of Waikato), can help us with our task at hand: classifying/scoring essays according to their level of English proficiency. Given that machine learning is quite appropriate for dealing with a large number of features and optimal at finding hidden patterns in data, we want to explore how suitable these algorithms are for dealing with the delicate and multivariate reality of second language proficiency. We are also interested in knowing if and how the outputs (results) of some classifiers (algorithms) might reflect common practice in Applied Linguistics. Finally, we would like to know whether there might be significant differences in how human raters differ from the factors we identify in terms of the accuracy of their classification.

Once we have identified the factors that determine essay grades, we shall be in a position to focus work on AES to automating the determination of those factors. We envision our work as contributing to this research line. In the following section we review work in AES to suggest how the present chapter might contribute to it.

2. Literature review

Automated Essay Scoring has been making substantial progress since its incipience, dated to the 1960s and the work of Page and his Project Essay Grading (PEG) system (Page, 1966). Contemporary systems make use of different techniques and frameworks in order to arrive at a classification for a given writing sample. The number of essays used to train the various systems also varies. We discuss here some of the main AES systems currently in use.

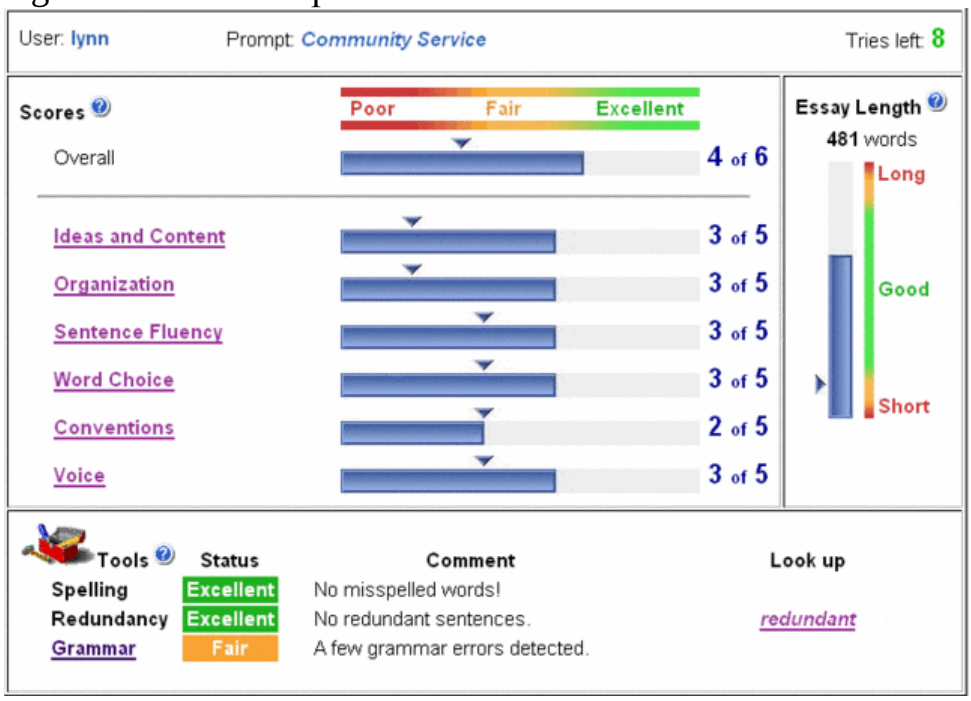
Page (1966), the developer of the PEG system, defines what he calls *trins* and *proxes*. *Trins* are intrinsic variables such as punctuation, fluency, grammar, vocabulary range, etc. As Page explains, these intrinsic variables cannot, however, be directly measured in an essay and must therefore be approximated by means of other measures, which he calls *proxes*. Fluency, for example, is measured through the prox "number of words" (Page, 1994). The main idea of the PEG system is quite similar to the one we have employed in our research. The system obtains as input a training set containing a large number of essays each with the values for the chosen proxes already assigned and a score for the overall writing quality. The system, by using regression analysis, arrives at the optimal weight for each of the proxes. For future ungraded essays, the system extracts from them the values for the same proxes used in the training phase and reaches a decision with regard to the level of the essay. The score generated is essentially a prediction of the grade that a teacher would have given for that specific essay (Rudner & Gagne, 2001).

Intelligent Essay Assessor™ (IEA) is an essay scoring system developed by Pearson Knowledge Analysis Technologies (PKT), which can provide analytic and trait scores, as well as holistic scores indicating the overall quality of the essay (PKT, 2011). This means that IEA can be used not only as an essay scoring system, but also in informative feedback, where students/test takers can identify the areas in which they are stronger and those that they need to work on (Figure 1). In addition to being able to analyze the more formal aspects of language, IEA also examines the quality of the content of essays, since it uses Latent Semantic Analysis (LSA), which according to Landauer et al (1998) is "a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (p. 259).

IEA is trained on a domain-specific set of essays (100-200 essays) that have been previously scored by expert human raters. When fed new essays whose score is unknown, IEA basically compares through LSA how similar the new essay is to the ones it has been trained on. If the new essay shows the highest similarity (both in terms of content and formal aspects, since these are not separate in LSA) to those essays in the corpus that have been scored a level 3, for example, that will be the score that IEA will output for the new essay in question. Through LSA, IEA is able to take into account not only formal linguistic features of the essays, but also deals with semantics, by representing each essay as a

multidimensional vector. According to the PTK website[5], the correlation between IEA and human graders is "as high or higher than that between two independent human raters" (PKT, 2011).

Figure 1. IEA™ sample feedback screen



Another quite well-known AES system is E-rater, developed by Education Testing Services (ETS) and employed in the TOEFL (Test of English as a Foreign Language) exam. The current version of E-rater is based on over 15 years of research on Natural Language Processing at ETS and takes several features into account when holistically scoring a writing piece (ETS, 2011). According to ETS, some of the features used by E-rater are: content analysis based on vocabulary measures, lexical complexity, proportion of grammar errors, proportion of usage errors, proportion of mechanical errors, organization and development scores, idiomatic phraseology and others (ETS, 2011). From these features, we see that E-rater goes beyond looking at only surface features and also examines organizational and developmental features, which makes it more suitable for use in a higher-stakes test such as the TOEFL than a system like PEG[6].

Many other systems have been developed, such as ETS1, Criterion[7], IntelliMetric[8] and Betsy[9], to mention a few. These systems vary considerably in

[5] http://www.pearsonpte.com/PTEAcademic
[6] http://www.measurementinc.com/News/PEG
[7] https://criterion.ets.org/
[8] http://www.vantagelearning.com/products/intellimetric/
[9] http://echo.edres.org:8080/betsy/

their approaches and methods for essay scoring. In 1996, Page makes a distinction between automated essay scoring systems that focus primarily on content (related to what is actually said) and those focusing primarily on style (surface features, related to how things are said) (as cited in Valenti, Neri & Cucchiarelli, 2003). Intelligent Essay Assessor [10], ETS1 and E-rater [11] are examples of the former type, while PEG and Betsy (a Bayesian system) are examples of the latter.

Rather than approach the problem of classifying texts according to grades directly, our strategy is to first attempt to isolate factors that correlate highly with test grades, in order to focus on automating these in a second step. We attempt thus to " divide and conquer", focusing here on dividing the problem into more manageable subproblems.

In our study we will primarily look at surface features, which include nonlinguistic (such as total number of words) and linguistic features (such as total number of grammatical errors), mainly because the texts are very short (about 150 words) and written by non-advanced learners of English (Dutch high school students).

3. Research context

In order to train a machine learning system, a corpus of holistically scored essays needs to be collected, so that it can be used as training data for the system. Since we are attempting to detect factors which are influential with respect to the holistically assigned grade, the training essays need to be annotated, meaning that we need to have a specific number of features that we look at in each essay and then record the value for each of those features (such as grammar mistakes, number of words, percentage of verbs in the present tense, etc).

The corpus we have used in our research comes from the OTTO project, which was meant to measure the effect of bilingual education in the Netherlands (Verspoor et al, 2010) and is the same as used by Verspoor et al (to appear). To control for scholastic aptitude and L1 background, only Dutch students from VWO[12] schools (a high academic Middle School program in the Netherlands) were chosen as subjects. In total, there were 481 students from 6 different VWO schools in their 1st (12 to 13 years old) or 3rd year (14 to 15 years old) of secondary education. To allow for a range of proficiency levels, the students were enrolled in either a regular program with 2 or 3 hours of English instructions per week or in a semi-immersion program with 15 hours of instruction in English per week. The 1st year students were asked to write about their new school and the 3rd year students were asked to write about their previous vacation. The word

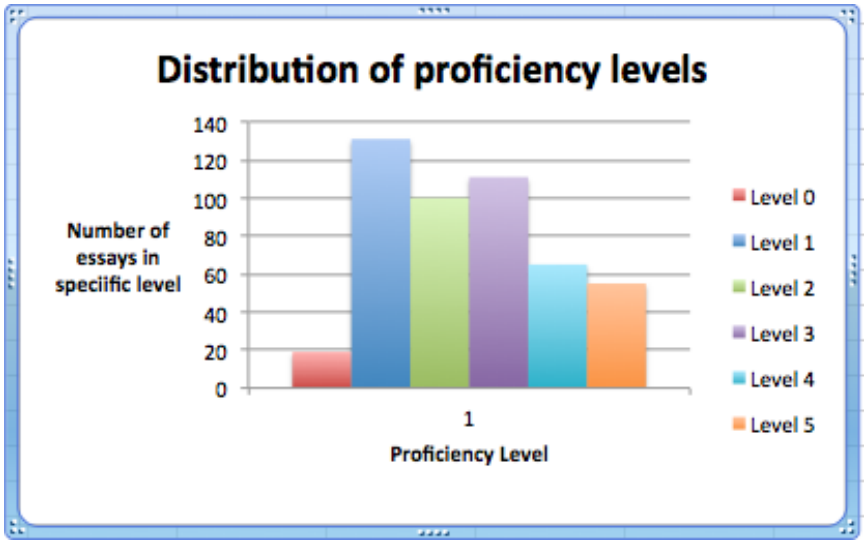
[10] http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf
[11] http://www.ets.org/erater/about
[12] VWO stands for *Voorbereidend Wetenschappelijk Onderwijs*, literally, Preparatory Scientific Education.

limit for the essays was approximately 200 words. The writing samples were assessed on general language proficiency and human raters gave each essay a holistic proficiency score between 0 and 5, with 0 indicating the level assigned to those essays with the most basic language complexity and 5 indicating those with the most complex language, out of the essays analyzed. As Burstein & Chodorow (2010) put it, "for holistic scoring, a reader (human or computer) assigns a single numerical score to the quality of writing in an essay" (p.529). In order to ensure a high level of inter-rater reliability, the entire scoring procedure was carefully controlled. There were 8 scorers, all of whom were experienced ESL teachers (with 3 of them being native speakers of English). After long and detailed discussions, followed by tentative scoring of a subset containing 100 essays, assessment criteria were established for the subsequent scoring of essays. Two groups of 4 ESL raters were formed and each essay was scored by one of the groups, with the score of the majority (3 out of 4) being taken to be the final score of the essay. If a majority vote could not be reached and subsequent discussion between the members of that group did not solve the issue, then the members of the other group were consulted in order to settle on the final holistic score for the essay. In all, 481 essays were scored. As we will see further ahead, the size of this set is good enough for training a scoring system and some of the more established essay scoring systems available actually use a smaller set than we do in our work. In Figure 2, we can find the distribution of the levels among the essays we have used.

Verspoor et al (to appear) coded each writing sample for features (variables) drawn both from the Applied Linguistics literature and from their own observations during the scoring of the essays. The features cover several levels of linguistic analysis, such as syntactic, lexical, mechanical, and others. Some of the features, such as range of vocabulary, sentence length, accuracy (no errors), type-token ratio (TTR), chunks, and number of dependent clauses, for example, are established features in the literature and have been used in several studies to measure the complexity of a written sample. Other features, such as specific types of errors and frequency bands for the word types were chosen in order to do a more fine-grained analysis of language.

In our study, we first investigated which machine learning algorithms found in WEKA, when trained on this corpus of variables, can achieve results which would allow them to be used in a future automated essay scoring system for a low-stakes test. In order to arrive at the optimal system, we experimented with decreasing the number of features available to the algorithms (through feature selection) and also discretizing the values (that is, using interval ranges instead of raw values for the features).

Figure 2. Distribution of the levels (0-5) in our data



Finally, we wanted to know how the results of the best trained system out of those analyzed might compare to the results seen when humans raters score the essays and how our results might reflect common practices in second language research.

4. Methodology

In machine learning, a common method of assessing the classification performance of a system is by doing what is called *ten-fold cross validation*. This method basically involves dividing the available data set into ten parts, with nine tenths serving as the training set and the one tenth which is left serving as the test set. This process is done 10 times (Jurafsky & Martin, 2009). Throughout our study, we have made use of the ten-fold cross validation method in order to assess the quality of the classification models. We have experimented with 2 different scenarios:

Scenario 1: All 81 features and their respective values are made available to the algorithms. We analyze their results in terms of absolute accuracy (assigning to an essay exactly the level it has been assigned by the human raters) and adjacent agreement (giving some credit for adjacent classifications as well).

Adjacent agreement is a looser measure of success when predicting classifications such as grades, where the classes are ordered. If we are dealing with a scale of 0-5 in terms of proficiency level, classifying a level 4 essay as level 5 or level 3 is certainly more desirable than classifying this same level 4 essay as level 1, for example. Therefore, we cannot take only absolute accuracy into account. In addition, human raters themselves quite often disagree on the exact level of a given essay, but tend to assign adjacent levels to the same essay.

At least, this is the desirable situation when raters have been as well trained as ours.

Scenario 2: We perform feature selection in order to find a subset of features that correlate highly with proficiency level and also submit the values of those features to discretization, before training three of our main systems. It is a known fact that obtaining comparable results by using fewer features is a gain in knowledge, given that it makes the model simpler, more elegant and easier to be implemented. Using every feature in order to build a classifier might also be seen as overkill. The question is simple: if we can achieve the same (or possibly even higher) accuracy in a system by using fewer features, why should we use all of them? It takes processing power and engineering/programming work in order for an automatic system to extract the values for each feature and if many of the features do not lead to an improvement in classification accuracy, it does not make much sense to insist on using them if our sole task is classification. In addition, by using too many features we might be missing some interesting patterns in our data. By discretizing numerical data we are able to build models faster, since numerical values do not have to be sorted over and over again, thus improving performance time of the system. On the other hand, discretizing values leads to a less fine-grained and transparent analysis, since we group together a continuum of values that might have individual significance for classification.

Finally, once we have analyzed our 2 scenarios and arrived at the optimal classification model out of those we have experimented with, we looked at how the model might be said to meet the gold standard and thus show results which are similar to those recorded when human raters grade the essays. For this, we needed not only classification accuracies and adjacent agreement, but also a new experiment in which we compare the results of the system and of the original human raters with those of a second and independent group of trained raters. In the next section, we report on the results of our experiments.

5. Results

5.1. Scenario 1

The accuracy of the 11 classifiers used in Scenario 1 (before feature selection and value discretization) is shown in Table 1 below. We would like to draw the reader's attention to the fact that the baseline classification accuracy for our data would be 27%, which is the result of dividing the number of essays belonging to the most common level (level 1 = 131 essays) by the total amount of essays in our corpus (481 essays).

Table 1. Accuracies (percentage of correct classification) of the 11 different classifiers, before feature selection and discretization

Classifier	**Ten-fold cross validation results (absolute accuracy in percentage)**	**Weighted scores (Cor = 3, Adj = 1, Inc = 0)**
LMT	58.09	1013
Functional Tree	56.07	980
Random Forest	53.97	1001
LAD Tree	53.49	973
Naïve Bayes	52.50	962
Simple Cart	52.10	949
Rep Tree	51.36	948
C4.5 (J48)	50.53	843
BF Tree	49.90	908
NB Tree	45.70	892
Decision Stump	40.73	762

As noticed in Table 1, Logistic Model Tree is the machine learning algorithm that not only manages to build the best classification model in Scenario 1, when taking only absolute accuracy into account, but also the model that scores the highest when adjacent classifications are taken into account as well.

5.2. Scenario 2

As we have just seen, LMT is the classifier that performs the best for our task when all 81 features are made available to the classifiers, both in terms of absolute accuracy and adjacent agreement. We now need to know whether doing feature selection and data discretization increases the accuracy of our systems and, if so, which of the classifiers performs the best.

By performing feature selection on our data, we arrive at a subset of 8 features (to be discussed in more detail in Section 6 of this chapter), which together, provide the optimal classification accuracy for the systems. The removal of any of these features from the subset leads to a decrease in classification accuracy. In Table 2 below we can see what those 8 features are in ascending order of how much they correlate with proficiency level, with feature 1 being the feature that correlates the highest with proficiency level.

Table 2. Feature selection based on the Infogain + Ranker method in WEKA

Rank of feature	Feature
1	Number of lexical types
2	Number of correct chunks
3	Number of correct + incorrect chunks
4	Percentage of no dependent clauses
5	Percentage of verbs in Present Tense
6	Percentage of errors in verb form
7	Percentage of lexical errors
8	Total number of errors

We have selected three of our classifiers in order to see the effect of feature selection and data discretization on their accuracy: LMT (our best classifier so far), Naïve Bayes (the only Bayesian classifier we have experimented with[13]) and C4.5 (arguably the most common benchmark in machine learning). We can find in Figure 3 below the results of feature selection and data discretization on these 3 classifiers.

Figure 3. C4.5, LMT and NB accuracies after pre-processing of data

Classifier	Previous accuracy (no pre-processing)	Accuracy (discretization only)	Accuracy (attribute selection only)	Accuracy (attribute selection + discretization)	Accuracy (discr. + attr.sel)
C4.5	50.53%	55.23%	52.93%	58.70%	59.53%
LMT	58.09%	62.29%	60.67%	62.58%	62.27%
Naïve B.	52.50%	60.73%	55.16%	59.09%	60.82%

As we can see in Figure 3, all 3 classifiers benefit from feature selection and discretization. However, the best result is achieved by LMT, when first attribute selection is performed, followed by the discretization of the values of the 8 features in the subset. Therefore, the best absolute frequency we have managed to achieve for our essay classification task is 62.58%.

In terms of adjacent classification, the optimal version of LMT (just discussed) achieves good results. In Figure 4 below, we see the adjacent agreement of LMT (classifying an essay as either its original level or an adjacent one):

[13] All other 10 classifiers are Decision Tree classifiers.

Figure 4. Adjacent agreement of LMT per level

	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Adjacent agreement	100%	98%	96%	94%	98%	94%

As can be noticed, whenever LMT does not assign the exact correct level for an essay, it assigns an adjacent level in the great majority of cases, which is exactly what one wants for our low-stakes AES task.

Finally, to see how LMT compares to human raters, we randomly selected 30 essays out of our 481 essays and asked a second and independent group of trained graders to score them. From Figure 5 we may conclude that LMT achieves a high correlation[14] with the second group, which is quite similar to the one observed for the original group of scorers.

Figure 5. Correlation coefficients in 2 conditions

	Human Raters group 2
Human Raters group 1	0.84
Logistic Model Tree (LMT)	0.87

Taking the results shown in this section, we can say that LMT has the potential to be a high-performing essay scoring model once the 8 features used can be automatically extracted from essays.

6. Discussion

Logistic Model Tree (LMT), when trained on 8 discretized features extracted from hundreds of essays, achieves results similar to those seen in human scorers. However, we cannot call LMT an AES system, since at this point it is not embedded in a system that extracts the 8 features and their corresponding values automatically.

The features identified show overlap to some extent with what the previous study by Verspoor et al (to appear) found using traditional statistical analyses, but some are surprising because they have thus far not been commonly associated with proficiency levels and were, therefore, not even considered in the previous

[14] We have used Pearson correlation in our calculation.

study. For example, the number of lexical types per text was not considered relevant in the previous study as the field of Applied Linguistics usually works with type-token ratios. *Types* refers to the raw number of types found in the text, not adjusted for text length. *Chunks* (also known as formulaic sequences) are target-like combinations of two or more words such as compounds, phrasal verbs, prepositional phrases and so on. The field of applied linguistics has long recognized a link between chunks and proficiency, but the previous study and this one are the first to show such a direct connection between chunks and proficiency level. The fact that incorrect chunks, which represent non-target attempts at formulaic language, are a strong predictor is unexpected and was, therefore, not considered to be relevant in the previous study. The remainder of the features was also identified in the previous study, but not in as much detail as in the present one. The percentage of "no dependent clauses" refers to the relative number of simple sentences in a text, a commonly found indicator of proficiency level. The percentage of present tense refers to the use of a simple present tense as opposed to the use of a past tense or of a modal, passive, progressive or perfect. The present tense is commonly recognized as the tense that beginners will use, but the fact that it is such a strong predictor is interesting. The last three features are not completely surprising as the number of errors has often been linked to proficiency level, but what is unexpected is that we do see that both different types of errors--verb form errors and lexical errors--and the total number of errors play such a strong role.

Now the question is to what extent these 8 features that correlate the most with proficiency level lend themselves differently to automation. Four of the features should pose no major problem and can be somewhat easily automated: *number of types*, *percentage of no dependent clauses*, *percentage of verbs in the present tense* and *percentage of errors in verb form*. The other four features are much more difficult to implement, given their intrinsic complexity: *number of correct chunks*, *number of correct and incorrect chunks*, *percentage of lexical errors* and *total number of errors*.

A few lines of code in any of the major programming languages for text analysis (such as Python) can extract the number of *types* in an essay. The amount of subordination has for a long time been used in the SLA literature to represent the syntactic complexity of texts (Michel et al., 2007). There are already systems available that are able to identify the number of clauses and *dependent clauses* in a sentence. One such system is the one developed by Xiaofei Lu (2010), called *L2 syntactic Complexity Analyzer*. The *percentage of verbs in the present tense* can be extracted by running each essay through a parser and morphological analyzer (these are both computational linguistics tools). Finally, the *percentage of errors in verb forms* can be identified in part by running the essays through a parser and for each verb checking whether the verb form can be found in a pre-determined database of existent verb forms in English,

for example. Errors due to inappropriate usage are more difficult to detect automatically.

Automatically extracting the other four features is a much more difficult task, especially due to the fact that each one of them contains various subtypes. In the *number of correct chunks* feature, for example, we find collocations, phrasal verbs, verb-preposition combinations (such as "depend on"), etc. Stefan Evert (2009), in an article entitled "Corpora and Collocation", summarizes a number of statistical methods that can be used for extracting collocations[15]. However, the author does not put them to the test, so we cannot be sure how accurate and appropriate they would be. We believe, however, that perhaps an n-gram based approach, in which we calculate the probability of unigrams, bigrams, trigrams, etc found in the essay based on a corpus of native English might be an alternative approach to determining unique word combinations.

The LMT model we have trained does have limitations, naturally. Firstly, since it only deals with surface features (no analysis of meaning/semantics is carried out), it is not appropriate for higher-stakes testing. Secondly, since the model has been trained by taking various features into account which might be typical of Dutch learners of English (such as some types of lexical and syntactical errors), performance of the system on essays written by learners of different L1s might show different results. Only future research will be able to show how the system would fair in such cases. Finally, there might be an effect of the prompts for the essays (the essay topics) in the features that were seen to correlate the most with proficiency level.

7. Conclusion

In the previous sections of this chapter, we have seen that machine learning techniques and algorithms can be of great use towards the task of identifying features that may lead to automated essay scoring. Machine learning can not only help with identifying a subset of features that correlate the most with the construct at hand (that is, the target class or proficiency level in our case) and therefore enhance the power and simplicity of the system by using only those features in classification, but also find patterns in the data, which can subsequently be used to classify new samples. Different machine learning algorithms make use of different strategies in order to arrive at their most optimal classification and show different classification accuracies. The Logistic Model Tree (LMT), which employs logistic regression to arrive at the most optimal classificatory function for each possible class (each of the proficiency levels the essays could belong to) manages to meet the gold standard by achieving the same classification accuracy observed in human scorers. In addition to LMT showing the same classification accuracy (both in terms of exact and adjacent

[15] Some of the methods include chi-squared, mutual information and Z scores.

classification) as human raters, the classification correlation coefficient observed between LMT and a group of human scorers equals the one recorded between 2 groups of trained human scorers. The 8 features found are interesting in themselves as they are not necessarily the ones that are commonly recognized in the Applied Linguistics literature and will contribute to insights into second language development. However, the 8 features found do not all easily lend themselves to automatic scoring. We hope though that once the subset of 8 features that have been used to build the LMT classification model can be extracted automatically or alternative features have been found, LMT has the potential to be used for the task of essay scoring, optimizing the scoring time, increasing the fairness of the scoring process and decreasing the need for human labor.

References

Educational Testing Services (ETS). (2011). Retrieved September 28, 2011, from http://www.ets.org/erater/how/

Burstein, J. & Chodorow, M. (2010). Progress and New Directions in Technology for Automated Essay Evaluation. In Kaplan, Robert.B (Ed.), Oxford Handbook of Applied Linguistics (pp. 529-538). Oxford University Press, 2010.

Jurafsky, D. & Martin J.H. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semanctic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-140

Lu, Xiaofei (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4): 474-496

Michel, M. C., F.Kuiken, & I.Vedder (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. International Review of Applied Linguistics in Language Teaching 45: 241-59.

N. Landwehr, M. Hall, and E. Frank. Logistic model trees. Machine Learning, 59(1- 2):161–205, 2005.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics, 30, 555-578.

Page, E. B. (1966). Grading essays by computer: Progress report. Notes from the 1966 Invitational Conference on Testing Problems, 87-100.

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

Pearson Knowledge Technologies (PKT). (2011). Retrieved September 28, 2011, from http://www.knowledge-technologies.com/papers/IEA_FAQ.html

Rudner, L. & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer (ERIC Digest number ED 458 290). Retrieved September 28, 2011, from http://www.eric.ed.gov/PDFS/ED458290.pdf

Stefan Evert. 2009. Corpora and Collocations. In: Anke Lüdeling and Merja Kytö, editors, Corpus Linguistics. An International Handbook, vol. 2. Mouton de Gruyter, Berlin: 1212-1248.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. Journal of Information Technology Education, 2, 319–330.

Verspoor, M.H., Schuitemaker-King J., Van Rein, E.M.J., De Bot, K., & Edelenbos, P. (2010). Tweetalig onderwijs: vormgeving en prestaties. Onderzoeksrapportage. (www.europeesplatform.nl/sf.mcgi?3847)

Verspoor, M., Schmid, M.S., & Xu, X (to appear). A dynamic usage based perspective on L2 writing development. *Journal of Second Language Writing.*