Automated Essay Scoring for Swedish

Robert Östling
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
robert@ling.su.se

Andre Smolentzov
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
asmolentzov@gmail.com

Björn Tyrefors Hinnerich
Department of Economics
Stockholm University
SE-106 91 Stockholm
bjorn.hinnerich@ne.su.se

Erik Höglin
National Institute of Economic Research
Kungsgatan 12-14
103 62 Stockholm
erik.hoglin@konj.se

Abstract

We present the first system developed for automated grading of high school essays written in Swedish. The system uses standard text quality indicators and is able to compare vocabulary and grammar to large reference corpora of blog posts and newspaper articles. The system is evaluated on a corpus of 1 702 essays, each graded independently by the student's own teacher and also in a blind re-grading process by another teacher. We show that our system's performance is fair, given the low agreement between the two human graders, and furthermore show how it could improve efficiency in a practical setting where one seeks to identify incorrectly graded essays.

1 Introduction

Automated Essay Scoring (AES) is the field of automatically assigning grades to student essays (Shermis and Burstein, 2003; Dikli, 2006).

Previous work on AES has primarily focused on English texts, and to the best of our knowledge no AES system for Swedish essays has been published. We exploit some peculiarities of the Swedish language, such as its compounding nature, to design new features for classification. We also use constructions in the shape of *hybrid n-grams* (Tsao and Wible, 2009) extracted from large corpora in the classification.

Earlier results from this work have been presented in the B.A. thesis of Smolentzov (2013), where further details can be found. Source code, a trained model as well as an on-line version of our tool are available from the website of the Department of Linguistics.[1] Due to legal restrictions, we are currently unable to publish the corpus of essays used for training the model and in our evaluation. While this is very regrettable, there are so far no suitable training corpora available for Swedish that are publicly available. We hope in the future to be able to produce an anonymized version of the corpus, to be shared with other researchers.

2 Data

We use a corpus of essays from the essay writing part of the Swedish high school national exams in Swedish.[2] These were collected using random sampling by Hinnerich et al. (2011), who had them digitized, anonymized, and re-graded by high school teachers experienced with grading the national exams. The essays were originally graded by the student's own teacher. In total, 1 702 essays have all the information we require: digitized text and the two grades. The size of the corpus is 1 116 819 tokens, or an average of 656 per essay. The essays have been automatically annotated with lemma and part of speech (PoS) information using Stagger (Östling, 2012).

There are four grades: IG (fail), G (pass), VG (pass with distinction) and MVG (excellent). Hinnerich et al. (2011) found that the agreement between the two human graders is rather low, and in the set of essays used in this study only 780 (45.8%) of the 1 702 essays received the same grade by both

[1] http://www.ling.su.se/aes

[2] Course *Svenska B*, fall 2005/spring 2006.

		Teacher				
		IG	G	VG	MVG	Sum
Blind grader	IG	74	147	50	5	276
	G	68	437	293	55	853
	VG	12	136	223	75	446
	MVG	1	25	55	46	127
	Sum	155	745	621	181	1 702

Table 1: Confusion matrix for the grades assigned by the students' own teachers, and during the blind re-grading process. In total, 780 essays (45.8%) are assigned the same grade. Linear weighted $\kappa = 0.276$

graders. In 148 cases (8.7%), the grade difference was more than one step.

In Table 1, we can clearly see that the blind graders' grades are generally lower. The disagreement is also more severe for the grades at the extremes of the scale.

It is important to note that the grading guidelines for the national exams do not focus exclusively on the quality of the language used, but rather on the ability of the student to produce a coherent and convincing argument, understanding and relating to other texts, or describing personal experiences. Some work has been carried out using high-level features in automated essay scoring. Miltsakaki and Kukich (2004) use some manual annotation to explore the role of coherence, and Attali and Burstein (2005) automatically analyze the overall structure of essays. Others take the contents of essays into account (Landauer et al., 2003), which is suitable for essay questions in non-language subjects.

We will, however, focus on form rather than content. One important reason for this is that our corpus of essays is spread out over 19 different topics (in several cases with as few as 20–30 essays each), where the type of text expected can vary considerably between topics.

3 Methods

We use a supervised machine learning approach, based on a Linear Discriminant Analysis classifier in the implementation of Pedregosa et al. (2011). Each essay is represented by a *feature vector*, whose contents we will describe in some detail in the following sections.

It is important to note that we are using *correlations* between grade and different features of the text, but the relationship between these features and the qualities of the essay on which the grade should be based may be complex. As a cautionary tale, we could mention that vocabulary related to cell phones was found to correlate strongly with essay grade. It turned out that poor students showed a strong preference for one of the given essay topics, which happened to center around cell phones. In the field of AES, it is particularly important to keep in mind that *correlation does not imply causation*.

3.1 Simple features

We use a number of features that may be directly measured from the text. These are presented below, roughly in decreasing order of correlation with essay grade. Most of the features have been discussed in previous literature on AES (Attali and Burstein, 2005), and specifically in the context of Swedish high school essays by Hultman and Westman (1977). Some further features that did not contribute much to grading accuracy were tried, but will be omitted from this discussion.

Text length Since the essays are composed in a classroom setting with a fixed amount of time allotted (five hours), a student's fluency in writing is directly mirrored in the length of an essay, which becomes the feature that most strongly correlates with grade. While one might want to exclude the length from consideration in the grading process, it is important to keep this correlation in mind since other measures may correlate with length, and therefore indirectly correlate with essay grade without contributing any new information.

Average word length The average number of letters per word also correlates with grade but only weakly with the length (in words). It does however correlate strongly with the distribution of parts of speech, primarily pronouns (which tend to be short) and nouns (which tend to be long, particularly since Swedish is a compounding language).

OVIX lexical diversity measure OVIX (Hultman, 1994) was in fact developed for the very purpose of analyzing lexical diversity in Swedish high school essays, and has been found to correlate

strongly with grade in this setting. At the same time, the measure is mostly independent of text length.

$$OVIX = \log n_{tokens} / \left(2 - \frac{\log n_{types}}{\log n_{tokens}}\right)$$

Part of speech distribution The relative frequencies of different parts of speech also correlate with essay grade, although more weakly so than the related measure of average word length.

3.2 Corpus-induced features

While the size our corpus of graded student essays is in the order of one million words, much larger amounts of Swedish text are available from different sources, such as opinion pieces, news articles, and blog posts. Due to the large amounts of text available, from tens of millions to several billions of words depending on the source, we can extract reliable statistics even about relatively rare language phenomena.

By comparing student essays to statistics gathered from different text types, we obtain new variables that often correlate strongly with essay grades.

PoS tag cross-entropy The average cross-entropy per token from a PoS trigram model (with simple additive smoothing) is used to model the similarity on a syntactic level. This includes both elements of style (e.g. frequent use of passive constructions) and mechanics (e.g. agreement errors). We use a corpus of news texts[3] to train the model.

Vocabulary cross-entropy With word frequency statistics from two different text sources, we compute the average cross-entropy per token given a unigram model, and use the difference between these values for the two models to indicate which type of text the present essay is most similar to. In our experiments, the two text sources are of equal size and consist of the news texts mentioned above, and a corpus of blog posts.

Hybrid n-gram cross-entropy We can generalize the vocabulary cross-entropy measure described above by using *hybrid n-grams* (Tsao and Wible, 2009) rather than single words. This allows for some patterns that are neither entirely grammatical nor entirely lexical to be used, complementing the two previous approaches. The same news and blog corpora as above are used.

[3]The corpus consists of ca 200 million words, crawled from the WWW editions of Dagens Nyheter and Svenska Dagbladet.

3.3 Language error features

Spelling errors We implemented a simple spell checker, using the SALDO lexicon (Borin and Forsberg, 2009) and statistics from a corpus of news text. On average, a misspelling was detected in 0.63% of all word tokens, or about four misspellings per essay. Manual inspection showed that the spell checker made some errors, so it is reasonable to assume that results could be improved somewhat using a more accurate tool.

Split compound errors Swedish is a compounding language, with noun compounding particularly frequent. It is a fairly common error among inexperienced writers to separate the segments of a compound word. We use word uni- and bigram statistics from a corpus of news texts to find instances of these errors in the essays. Only 0.10% of word tokens are found to be incorrectly split, or less than one instance per essay on average. As expected, there is a (weak) negative correlation between split compound frequency and grade, which seems to be due to a small number of poor essays with many such errors.

3.4 Evaluation measures

The simplest measure of overlap between two graders (either among humans, or between human(s) and machine) is the percentage of essays on which they agree about the grade. However, in our setting this is not so informative because there is a high chance of graders assigning the same grade by chance, and this probability varies between different pairs of graders.

This makes comparisons difficult, so we instead use Cohen's kappa value (Cohen, 1968), linearly weighted according to the numeric values of grades used by the Swedish school system: IG corresponds to 0 points, G to 10, VG to 15, and MVG to 20. A kappa value of 1 would indicate perfect agreement, while 0 would mean random agreement. The

Feature	Correlation
$n_{tokens}{}^{0.25}$	0.535
n_{tokens}	0.502
hybrid n-gram cross-entropy	0.363
vocabulary cross-entropy	0.361
average word length	0.307
OVIX	0.304
n_{long}/n_{tokens}	0.284
spelling errors	-0.257
PoS cross-entropy	0.216
split compound errors	-0.208

Table 2: Correlation between grade (average of two graders) and features. Interactions between features are not taken into account. Only features with Pearson coefficient $\rho > 0.2$ are included, all are highly significant.

weighted kappa value is computed as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}$$

where O_{ij} is the number of times annotator 1 assigned grade i and annotator 2 assigned grade j, while E_{ij} is the *expected* number of times for the same event, given that both annotators randomly assign grades according to a multinomial distribution. w_{ij} is the difference in score between grades i and j, according to the above.

4 Results

4.1 Feature-grade correlations

First, we look at the correlations between the human-assigned grades and individual features. Since a linear machine learning algorithm is used, we use the Pearson coefficient to measure linear dependence. Spearman's rank correlation coefficient gives similar results.

From Table 2 we can see that only ten of the features show a correlation above 0.2. There were statistically significant (but weak) correlations below this threshold, e.g. the ratios of different parts of speech, where the strongest correlations were $\rho = -0.192$ (pronouns) and $\rho = 0.177$ (prepositions).

4.2 Automated grading

Table 3 shows the performance of our system, using the leave-one-out evaluation method on all 1 702 essays, i.e. evaluating each essay using a model trained on all the other 1 701 essays. We see that the computer's grades are biased towards the most common grade (G, pass), but that overall accuracy is quite high (62.2%, $\kappa = 0.399$) compared to 58.4% ($\kappa = 0.249$) when using only the strongest feature (4th root of essay length), 54.9% when assigning the most common grade to all essays, or the 45.8% ($\kappa = 0.276$) agreement between the two human graders.

		Computer				
		IG	G	VG	MVG	Sum
Human avg.	IG	107	176	6	0	289
	G	61	752	110	11	934
	VG	2	225	189	17	433
	MVG	0	9	27	10	46
	Sum	170	1 162	332	38	1 702

Table 3: Confusion matrix for the grades assigned by the system, and the average (rounded down) of the two human graders. In total, 1 058 essays (62.2%) are assigned the same grade, $\kappa = 0.399$.

It is also encouraging to see that only 28 essays (1.6%) receive a grade by the computer that differs more than one grade from the human-assigned grade. The corresponding figure is 148 essays (8.7%) between the two humans.

When training and evaluating using only the grades of the blind grader, the agreement between computer and human was 57.6% ($\kappa = 0.369$), and only 53.6% ($\kappa = 0.345$) using the grades of the student's teacher. Both these figures are below the 62.2% ($\kappa = 0.399$) obtained when using the average grade, and the explanation closest at hand is that the features we model (partially) represent or correlate with the actual grading criteria of the exam.

Since the teachers are affected by various sources of bias (Hinnerich et al., 2011), a weaker correlation (mirrored by a lower κ) to any kind of "objective" measure would be expected. Similarly, using the average of two graders should decrease the large individual variance due to the difficult and partially subjective nature of the task, leading to a stronger correlation with relevant features of the text.

4.3 Re-grading

In 148 cases (8.7%) of our 1 702 essays, the grade assigned in the blind re-grading process differs by more than one step from the original grade, and we performed an experiment to see how efficiently these *highly deviant* grades could be identified. This scenario could arise within an organization responsible for evaluating the consistency in grading a national exam, where resources are insufficient for re-grading *all* essays manually. Given a training corpus of graded essays, our system could then be used to select candidates for further manual processing.

In other to evaluate the usefulness of this method, we let the system re-grade all essays based on the blind grades (using the leave-one-out method). In the cases when the system's grade differs by more than one step from the teacher's grade, we check whether the difference between the system's grade and that of the blind grader is less than between the two human graders. It turns out that we can correctly identify 43 (29.1%) of the 148 cases in this way, with only 91 essays (5.3% of the total) considered.

In a scenario where we have a large amount of essays but only the resources to manually re-grade a fraction of them, we can thus increase the ratio of highly deviant grades found from 8.7% ($148/1702$, by randomly choosing essays to re-grade) to 47% ($43/91$, by only re-grading those identified by our system).

5 Conclusions and future work

We have presented a system for automatic grading of Swedish high school essays, and demonstrated its usefulness in a practical setting of finding instances of incorrect grading. Novel aspects include features based on constructions induced using unsupervised methods, and on (language-specific) compounding errors.

It would be interesting to apply some of our methods to other languages and other data sets, for instance of second language learners. Since our system is quite general, all that would be needed to adapt it to another domain is a training corpus of graded essays. Adapting to another language would in addition require a PoS tagger and suitable unlabeled text corpora (e.g. of blogs and professional prose).

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments.

References

Yigal Attali and Jill Burstein. 2005. Automated essay scoring with e-rater® v.2.0. Technical report, Educational Testing Services.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Semire Dikli. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5.

Björn Tyrefors Hinnerich, Erik Höglin, and Magnus Johannesson. 2011. Are boys discriminated in swedish high schools? *Economics of Education Review*, 30:682–690.

Tor G. Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. LiberLäromedel.

Tor G. Hultman. 1994. Hur gick det med ovix? In *Språkbruk, grammatik och språkförändring. En festskrift till Ulf Teleman*, pages 55–64. Lund University.

Thomas K. Landauer, Darrell Laham, and Peter Foltz. 2003. Automatic essay assessment. *Assessment in Education*, 10:295–308.

E. Miltsakaki and K. Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10:25–55.

Robert Östling. 2012. Stagger: A modern POS tagger for Swedish. In *Proceedings of the Swedish Language Technology Conference (SLTC)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

M.D. Shermis and J. Burstein, editors. 2003. *Automated Essay Scoring: A Cross Disciplinary Perspective*. L. Erlbaum Associates.

André Smolentzov. 2013. *Automated Essay Scoring: Scoring Essays in Swedish.* Bachelor's thesis, Department of Linguistics, Stockholm University.

Nai-Lung Tsao and David Wible. 2009. A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 51–54, Stroudsburg, PA, USA. Association for Computational Linguistics.