A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese

Evelin Carvalho Freire de Amorim
Computer Science Department
Universidade Federal de Minas Gerais
Minas Gerais, Brazil
evelin.amorim@dcc.ufmg.br

Adriano Veloso
Computer Science Department
Universidade Federal de Minas Gerais
Minas Gerais, Brazil
adrianov@dcc.ufmg.br

Abstract

While several methods for automatic essay scoring (AES) for the English language have been proposed, systems for other languages are unusual. To this end, we propose in this paper a multi-aspect AES system for Brazilian Portuguese which we apply to a collection of essays, which human experts evaluated according to the five aspects defined by the Brazilian Government for the National High School Exam (ENEM). These aspects are skills that student must master and every skill is assessed separately from one another.

In addition to prediction, we also performed feature analysis for each aspect. The proposed AES system employs several features already used by AES systems for the English language. Our results show that predictions for some aspects performed well with the employed features, while predictions for other aspects performed poorly.

Furthermore, the detailed feature analysis we performed made it possible to note their independent impacts on each of the five aspects. Finally, aside from these contributions, our work reveals some challenges and directions for future research, related, for instance, to the fact that the ENEM has over eight million yearly enrollments.

1 Introduction

The goal of automatic essay scoring (AES) systems is to score a given essay. AES systems are relevant for educational institutions, since the human effort to evaluate essays is high and, and students need feedback to improve his or her writing skills. Besides these issues, almost every senior high school student in Brazil should write an essay to the National Exam of High School (ENEM), which Brazilian government uses to evaluate the quality of high schooler's education and that of their institution.

Although there are thousands of essays written to ENEM every year, to the best of our knowledge there is no AES system for Brazilian Portuguese (BP) language, or an analysis of features in a multi-aspect essay scoring system for BP. Each aspect is a skill that students must master as seniors in high school. Nonetheless, several AES systems have been proposed for the English Language. Attali and Burstein (Attali and Burstein, 2006) proposed an AES system, called e-rater, that employs general features of argumentative essays to scoring prediction. The main features used by e-rater are grouped into the following types: grammar, usage, mechanics, and style; organization and development; lexical complexity; and prompt-specific vocabulary usage. e-rater employs multiple regression, and it is task independent AES system, i.e., its score is independent of the given prompt.

Napoles and Callison-Burch (Napoles and Callison-Burch, 2015) employ linear regression to an AES system that intends to assign more uniform grades than multiple human evaluators. Similar to our task, Napoles and Callison-Burch propose the task of multi-aspect classification using five grading categories. However, the authors leave unexplained how each of their aspects is affected by their features. We think this is a significant contribution since students or professors can use features as a feedback for better understanding essays writing. Besides that, Napoles and Callison-Burch assume that more than one evaluator is available to train their model, which in the real world is not always the case.

Larkey (Larkey, 1998) proposed three models

Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 94–102, Valencia, Spain, April 3-7 2017.

that are based on text classification to score essays applying linear regression. However, Larkey strategy is task-dependent. Chen and He (Chen and He, 2013) also grouped features into four main types: lexical features; syntactical features; grammar and fluency features; content and prompt-specific features. Then, the authors proposed a rank-based algorithm that maximizes the agreement between human score and machine score. Zesh et al. (Zesch et al., 2015) developed a technique that adapts domain in an AES system. The authors tested their method in an English dataset and a German dataset. Chali and Hasan (Hasan, 2012) proposed an LSA-based method, that is task-dependent, and the goal was to establish a strategy to understand the inner meaning of texts. Beyond the English language, Kakkonen and Sutinen (Kakkonen and Sutinen, 2004) developed an AES system to the Finnish language also based on LSA algorithm.

Besides assigning grade score, other researches proposed to analyze argumentation strength of essays (Persing and Ng, 2015), discourse structure of essays (Song et al., 2015)(Stab and Gurevych, 2014) , and grammar correction in general (Rozovskaya and Roth, 2014)(Lee et al., 2014).

Our research is different from the previous research since we aim to answer the following questions:

1. How objective features behave in a multi-aspect automatic essay scoring system?
2. Which features are more relevant for each aspect?

In addition to this, we aim to pose some interesting questions for future research. Our essays present not only grades but also evaluator's comments about the aspects that are considered in the ENEM. During the exploration of evaluators comments, bias was observed in some evaluations, which we define as being when some human evaluator seems to disagree or agree with student's point of view, which can lead to an improper influence on the student's grade. The possibility of bias of human evaluations raises some questions.

1. Are some topics for essays more prone to result in biased evaluation?
2. Is it possible to detect if human evaluator is biased for or against a given student's point of view?
3. If it is possible to detect the bias of human evaluator, is it feasible to measure the quantitative affect on grades?
4. Is there any difference between the words in biased evaluations that agrees with student point of view and biased evaluations that disagrees with student's point of view?

In a nutshell, the availability of evaluator comments allows for a host of issues related to bias detection, quantification, and resolution, yet as far as we know these questions are still unanswered.

The paper is organized as follows. The second section details our dataset and the features we use. The third section explains the experiments we performed and the results of our experiments. The fourth section presents the main remarks about our research and the fifth section point to the future direction for our research.

2 Methodology

We propose a methodology that besides the usual features employed by popular AES methodologies (Attali and Burstein, 2006) (Chen and He, 2013) (Zesch et al., 2015), it also takes advantage of domain features. To test our proposed features, we used a dataset of nearly 1840 essays. Next sections describe our dataset and our features.

2.1 Dataset

Our dataset is composed of 1840 essays about 96 topics, which were crawled from UOL Essay Database website[1]. The average length in words are 300.51; the biggest essay has 1293 words, and the smallest essay has 49 words. Each essay is evaluated according to the following five aspects:

1. **Formal language**: Mastering of the formal Portuguese language.
2. **Understanding the task**: Understanding of essay prompt and application of concepts from different knowledge fields, to develop the theme in an argumentative dissertation format.
3. **Organization of information**: Selecting, connecting, organizing, and interpreting information, facts, opinions, arguments to advocate a point of view.

[1] http://educacao.uol.com.br/bancoderedacoes

Table 1: Score and corresponding levels

Score	Level
2.0	Satisfactory
1.5	Good
1.0	Regular
0.5	Weak
0.0	Unsatisfying

Table 2: Average Score for each aspect and final grade in UOL Dataset

Aspect	Average Score
Formal Language	1.1
Understanding the task	0.91
Organization of information	0.93
Knowing argumentation	0.83
Solution proposal	1.08
Final grade	4.86

4. **Knowing argumentation**: Demonstration of knowledge of linguistic mechanisms required to construct arguments.

5. **Solution proposal**: Formulation of a proposal to the problem presented, respecting human rights and considering socio-cultural diversity.

Each aspect is scored according to the scale of Table 1, and the final score is the sum of all aspects scores. Table 2 depicts the average score assign by humans for each aspect and final grade in our dataset.

Each essay is evaluated by only one human. Although this seems a disadvantage, we think that this is a real world dataset, since in most high schools only one teacher scores essay. Also, as we aim to detect the impact of features in each aspect, one evaluator per essay is enough.

2.2 Features

Features are divided into two main types, domain features that are related to ENEM exam or Brazilian Portuguese Language, and general features that are based on Attali and Burstein research (Attali and Burstein, 2006).

1. *Domain features:* ENEM exam doesn't allow the using of the first person pronouns and verbs. Therefore, we employ as features the number of first person pronouns and verbs and the number of first person pronouns and verbs per number of tokens. Also, we suggest as feature the number of *ênclise*, a Portuguese language structure, and the number of *ênclise* per number of tokens. *Ênclise* is unusual to BP spoken language, then if a student applies such concept in essay, probably he or she knows how to use formal language better. Also, the excessive number of demonstrative pronouns is condemned in written BP (Martins, 2000); then we use the number of demonstrative pronouns and the number of demonstrative pronouns per number of tokens.

2. *General:* Most of the general features are based on Attali and Burstein (Attali and Burstein, 2006) research, which presented ten features. However, due to lack of tools for Brazilian Portuguese and time constraints, we implemented only six features and adapted two features from the e-rater framework. Next, we detailed our feature implementation.

   - *Grammar and style:* Grammar was checked by CoGrOO (Kinoshita et al., 2006), which is a Brazilian add-on to Open Office Writer. Also, for spelling mistakes we use a Brazilian software [2]. Both features were also divided by the number of tokens in an essay; then we employed four features for grammar and spelling errors. To evaluate style in essays, we applied LanguageTool rules for Portuguese, but also we added some rules suggested by a Portuguese manual of writing (Martins, 2000)[3]. We employed the number of style errors and the number of style of errors per sentence as features.
   - *Syntactical features:* According to (Martins, 2000), in Portuguese Language, sentences longer than 70 characters are long sentences, and therefore are not recommended. We employ as a feature, the number of sentences longer than 70 characters.
   - *Organization and development:* There

[2] https://github.com/giullianomorroni/JCorretorOrtografico

[3] Rules can be examined in `https://goo.gl/F32hcC`

are no tools to evaluate organization and development in Portuguese language, then we collected discourse markers in a Brazilian Portuguese grammar (Jubran and Koch, 2006). Discourse markers are linguistic units that establish connections between sentences to build coherent and knit discourse. We employed as features the number of discourse markers and the number of discourse markers per sentence.

- *Lexical complexity:* To evaluate lexical complexity, we used four features. The first feature is Portuguese version of Flesh score (Martins et al., 1996); the second feature is average word length, which length is the number of syllables; the third feature is the number of tokens in an essay; the fourth feature is the number of different words in an essay.
- *Prompt-specific vocabulary usage:* It is desirable to employ concepts from the prompt in the essay, therefore for each essay we compute cosine similarity between prompt and essay. In this case, the prompt is a frequency vector of words, and the essay is also a frequency vector of words, which are from the prompt vocabulary. We decided for this strategy since, unlike other works, our dataset comprises many different topics, each with few essays. Then, we think that build a vocabulary for each domain it is not helpful.

3 Experiments

We performed two types of experiments: one evaluating the performance of grade prediction for each aspect and other evaluating the role of each feature in grade prediction task. Feature analysis is of particular importance for this task since computer evaluation of an essay is different from a human analysis. Therefore, explore which variable is important for which aspect is crucial for the development of our research.

3.1 Prediction Analysis

Besides ASAP challenge at Kaggle[4], several works employ **quadratic weighted kappa** as the evaluation metric (Zesch et al., 2015)(Chen and He, 2013)(Attali and Burstein, 2006), which aims to measure agreement between human evaluation and machine scoring. When the value of kappa is closer to 1, the higher the agreement between evaluators, and when the value of kappa is closer to 0, the lower the agreement between evaluators.

[4]https://www.kaggle.com/c/asap-aes/details/evaluation

Table 3: List of Features grouped into domain type and general type

Group	Feature
Domain	#first person of singular of verbs and pronouns
	#first person of singular of verbs and pronouns / #tokens
	#demonstrative pronouns
	#demonstrative pronouns / #tokens
	#enclise
	#enclise / #tokens
General	#sentences longer than 70 characters
	#grammar errors
	#grammar errors / #token
	#spelling errors
	#spelling errors / #token
	#style errors / #sentences
	#discourse markers
	#discourse markers / #sentence
	Flesh score
	Average word length (syllables)
	#tokens
	similarity with prompt
	#different words

First, we compute a matrix of weights (Equation 1) that are based on the difference between human evaluation and machine scoring.

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (1)$$

The second step calculates a histogram matrix called O, where $O_{i,j}$ is the number of essays that receive grade $i \in N$ by a human evaluator and a grade $j \in N$ by a machine evaluator. After that, we built another matrix E of expected ratings, which is the outer product between each rater's histogram vector of ratings. Finally, we employ O, E, and w to compute the quadratic weighted kappa using Equation 2.

Table 4: Kappa values for each grade aspect

Grade Type	Kappa
Final Grade	0.3673
Formal Language	0.3147
Understanding the task	0.2678
Organization of Information	0.2305
Knowing argumentation	0.2704
Solution proposal	0.1393

Table 5: Kappa values for each grade aspect after oversampling (full and general feature set)

Grade Type	Full	General
Final Grade	0.4245	0.4131
Formal Language	0.3351	0.3249
Understanding the task	0.1817	0.1822
Organization of Information	0.2728	0.2679
Knowing argumentation	0.2668	0.2484
Solution proposal	0.1542	0.1430

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad (2)$$

A simple regression is applied to predict the final grade of essays, and each of other five aspects. Also, a simple oversampling strategy is applied since grade distribution is unbalanced (Figure 1).

In the first step of oversampling strategy, it searches by the class G_{max} that holds the largest number of instances. Then the strategy randomly selects instances from every class $G \neq G_{max}$ and replicates such instances into training datasets, until the size of every class $G \neq G_{max}$ be equal the size of G_{max}.

Table 4 describes the results using quadratic weighted kappa before the oversampling. We executed cross-validation five times and compute the average of kappas of all experiments, for each aspect and final grade, to evaluate oversampling performance. Results after oversampling are described in Table 5.

Considering the lack of tools for processing the Portuguese language, and the limited performance of the few existing tools, the multi-aspect classification performed satisfactorily. However, some aspects performed poorly probably due to the subjectivity intrinsic to these aspects and objective variables probably can't capture all the subjectivity.

3.2 Feature Analysis

Besides kappa results, we also performed an experiment that investigates the impact of each feature in each aspect and final grade. The experiments were performed removing each feature and measuring the resulting kappa. If removing a feature f lowers the resulting kappa, then that feature is relevant to the model of that aspect. According to this criterion, the lower the resulting kappa when removing f from the training model, the more important is f for this model. Table 3.2 describes the three features that most diminished kappa value and the three features that most increased kappa value. The **full** value in table present the result with the full set of features described earlier.

It is possible to observe that the most relevant features for the final grade are not necessarily a mix of relevant features from the aspects. For instance, vocabulary level is one of three most important features for the final grade, but, while not irrelevant, it is not in the top three for the aspects. To understand better the role of vocabulary level, we compute in our dataset average vocabulary level for the final grade, and, as expected, the higher the grade, the higher the number of different words in essays. Besides vocabulary level, lexical complexity seems to play a significant role to final grade, since three of the most important features to final grade prediction affect prediction.

Aspect *Understanding the task* presented the lowest kappa value between aspects. However, we can draw some conclusions from Table 3.2. For instance, Flesh score affected expressively kappa value. Also, we observe that current features are not enough for *Understanding the task* model, therefore we will implement new features related to this aspect.

Organization of information resulted in the second highest kappa value between aspects. As *similarity to prompt* was the most relevant features, we believe that similarity between semantic vectors, as proposed by Zesh et. al (Zesch et al., 2015), also can improve *Organization of Information* prediction. Another observation is the influence of style errors upon *Organization of Information* aspect. Perhaps this influence is because the definition of style we used is related to how the writer "present" the information, which can be redundancies or nonexistent language expressions.

With respect to the *Knowing argumentation* as-

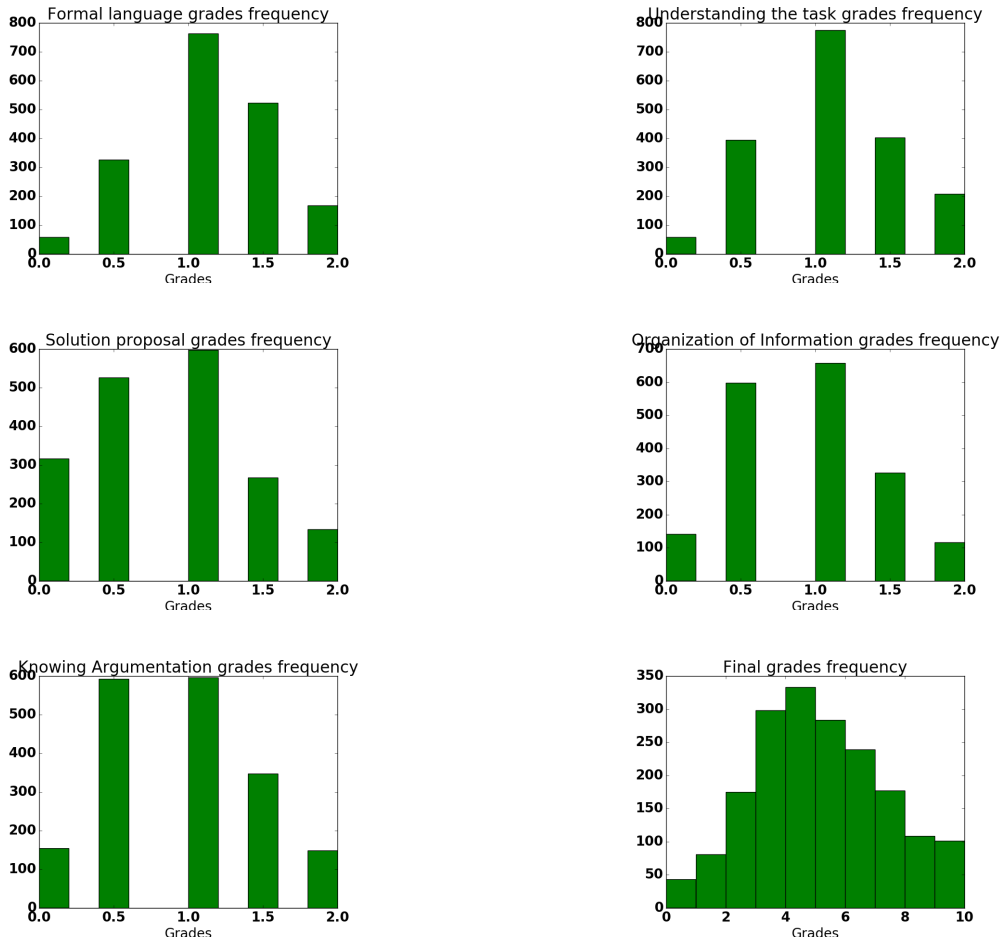

Figure 1: Distribution of grades in UOL dataset for each aspect and final grade

pect, we believe that style errors affected results for a similar reason that we mentioned in the analysis of *Organization of information* aspect. However, in regard this aspect we think that perhaps some argument features ((Stab and Gurevych, 2014), (Song et al., 2015)) will improve *Knowing argumentation* scoring prediction.

4 Conclusion

We proposed a multi-aspect automatic essay correction system for Brazilian Portuguese. Our primary goal is to evaluate if classical features for AES system for the English language performs well in a multi-aspect scenario, and assess which features are important for which aspect. In fact, after experiments, some features performed well for some aspects. Nonetheless, each aspect performed in a different way, which suggests that each aspect needs an own suitable model. Also, more specific features for some aspects probably will enhance subjective aspects.

Academic level, represented by Flesh score, is extremely relevant in most aspects. A possible reason for these results is because a high school student should present advanced skills, like grammar, spelling, argumentation, among others. Despite this feature in common, each aspect exhibits their singularity. Like enclise affecting *Understanding the task*, similarity with prompt influencing *Organization of information*, and discourse markers changing *Solution proposal*. Therefore, while some of the features enhance results for some aspects, these same features harm prediction for other aspects.

5 Future Directions

The following issues are directions we aim to pursue in our further research.

Analysis of evaluators comments. Our dataset comprises human evaluators comments. We intend to analyze these comments, which is of particular importance for argumentative essays since the opinion of human evaluators about a topic can affect grades. In a sample of 48 essays taken from our dataset, two linguists detected that 11 essays presented biased evaluation. Biased evaluation is a more serious issue if we will think about ENEM and other tests that are a relevant factor to many students. Some works were performed in bias language, but none of them analyzed bias on evaluations. Also, we can apply the same reasoning for other types of evaluations, like peer review of papers. Besides that, we would like to research how we can minimize bias on automatic scoring prediction.

Composite Classifier. A classifier to predict final grades employing predictions of the five aspects is a natural step in our research.

Adding new features to Brazilian Portuguese AES. There are more features to add to Brazilian Portuguese AES. Some of these features are: POS-tagging ratio; word length in characters; the number of commas, quotations or exclamation marks; average sentence length; average depth of syntactic trees; and topical overlap between adjacent sentences. Also, cohesion features like proposed by Song et al. (Song et al., 2015) can improve aspects like *Solution Proposal*, which probably demands sophisticated features.

6 Acknowledgements

We would like to thank Marcia and Luana, the two linguists that have been assisting us on bias research.

We thank the partial support given by the Brazilian National Institute of Science and Technology for the Web (grant MCT-CNPq 573871/2008-6), Project Models, Algorithms and Systems for the Web (grant FAPEMIG/PRONEX/MASWeb APQ-01400-14), and authors individual grants and scholarships from CNPq and CAPES.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, WA, USA.

Yllias Chali Sadid A Hasan. 2012. Automatically assessing free texts. In *24th International Conference on Computational Linguistics*, page 9, Bombay, India.

Clélia Jubran and Ingedore Koch. 2006. *Gramática do português culto falado no Brasil: construção do texto falado*, volume 1. UNICAMP.

Tuomo Kakkonen and Erkki Sutinen. 2004. Automatic assessment of the content of essays based on course materials. In *2nd International Conference on Information Technology: Research and Education*, pages 126–130, Semarang, Indonesia. IEEE.

Jorge Kinoshita, Lais N. Salvador, and Carlos E. D. Menezes. 2006. Cogroo: a brazilian-portuguese grammar checker based on cetenfolha. In *The fifth international conference on Language Resources and Evaluation (LREC)*, pages 2190–2193, Genova, Italy.

Leah S Larkey. 1998. Automatic Essay Grading Using Text Categorization Techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95, Melbourne, Australia.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, Yuen-Hsien Tseng, Li-Ping Chang, and Hsin-Hsi Chen. 2014. A sentence judgment system for grammatical error detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 67–70, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Teresa B. F. Martins, Claudete M. Ghiraldelo, Maria G. V. Nunes, and Osvaldo N. Oliveira Junior. 1996. *Readability formulas applied to textbooks in brazilian portuguese*. Instituto de Ciłncias Matemticas de So Carlos-USP, São Carlos, Brazil.

E. Martins. 2000. *Manual de redação e estilo*. O Estado de São Paulo.

Courtney Napoles and Chris Callison-Burch. 2015. Automatically scoring freshman writing: A preliminary investigation. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 254–263, Denver, CO, USA.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China.

Alla Rozovskaya and Dan Roth. 2014. Building a state-of-the-art grammatical error correction system. *Transactions of the Association for Computational Linguistics*, 2:419–434.

Wei Song, Ruiji Fu, Lizhen Liu, and Ting Liu. 2015. Discourse Element Identification in Student Essays based on Global and Local Cohesion. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2255–2261, Lisbon, Portugal.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Doha, Qatar.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-Independent Features for Automated Essay Grading. pages 224–232, Denver, CO, USA.

Table 6: Kappa results for Feature Analysis

Aspect	Feature Category	Feature Removed	Kappa
Final Grade	Most Relevant Features	Average word Length	0.3890
		Flesh Score	0.4010
		Vocabulary Level	0.4059
	Least Relevant Features	Discourse markers per #Sentence	0.4259
		Count of first Person	0.4262
		Count first Person per #Sentence	**0.4320**
		Full feature set	0.4245
Understanding the Task	Most Relevant Features	Flesh Score	0.1452
		#enclise / #sentences	0.1655
		#spelling errors	0.1655
	Least Relevant Features	#grammar errors	0.1868
		#style errors / #sentences	0.1878
		#first person use / # sentences	**0.1885**
		Full feature set	0.1817
Organization of Information	Most Relevant Features	Similarity with prompt	0.2496
		Average word length	0.2581
		#style errors / #sentences	0.2605
	Least Relevant Features	#long sentences	0.2788
		#demonstrative pronoun / # sentence	0.2799
		#first person use / #sentence	**0.2817**
		Full feature set	0.2728
Knowing Argumentation	Most Relevant Features	#spelling errors / #tokens	0.2438
		#style errors / #sentences	0.2441
		Flesh Score	0.2456
	Least Relevant Features	#enclise / #sentences	0.2773
		Average Word Length	0.2784
		#grammar errors	**0.2849**
		Full feature set	0.2668
Solution Proposal	Most Relevant Features	Average word length	0.1048
		Flesh score	0.1192
		#discourse markers	0.1240
	Least Relevant Features	#grammar errors / #Tokens	0.1586
		#tokens	0.1593
		#first person use	**0.1655**
		Full feature set	0.1655
Formal Language	Most Relevant Features	Flesh Score	0.3060
		#grammar errors / #tokens	0.3138
		#spelling mistakes	0.3248
	Least Relevant Features	#long sentences	0.3396
		#discourse markers	0.3396
		#demonstrative pronouns	**0.3429**
		Full feature set	0.3351