Using Analytic Scoring Rubrics in the Automatic Assessment of College-Level Summary Writing Tasks in L2

Tamara Sladoljev-Agejev[1] and **Jan Šnajder**[2]
[1] University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
[2] University of Zagreb, Faculty of Electrical Engineering and Computing,
Text Analysis and Knowledge Engineering Lab, Zagreb, Croatia
tagejev@efzg.hr, jan.snajder@fer.hr

Abstract

Assessing summaries is a demanding, yet useful task which provides valuable information on language competence, especially for second language learners. We consider automated scoring of college-level summary writing task in English as a second language (EL2). We adopt the Reading-for-Understanding (RU) cognitive framework, extended with the Reading-to-Write (RW) element, and use analytic scoring with six rubrics covering content and writing quality. We show that regression models with reference-based and linguistic features considerably outperform the baselines across all the rubrics. Moreover, we find interesting correlations between summary features and analytic rubrics, revealing the links between the RU and RW constructs.

1 Introduction

Writing summaries is a complex skill which relies on reading comprehension and the ability to convey the information contained in the source text(s). This makes summaries an important skill to develop for academic or professional purposes. Summary writing skills may therefore be tested in a recruitment process, or during admissions to universities, which may be particularly challenging for L2 writers who may still be struggling with lower levels of language competence such as grammar or vocabulary. This is why summary writing is sometimes used together with essays to assess university-level abilities in L2.

However, assessing L2 summaries is highly demanding, especially if analytic rubrics are involved, as they require raters' expertise and much concentration when assessing language proficiency at various levels (e.g., lexis, syntax, discourse). Moreover, unlike in essays, in summaries raters are expected to put additional effort into checking for accuracy, relevance, completeness, and coherence of the summary against the source text. Automated scoring is thus of considerable importance to enhance assessment of summaries, especially in the context of higher education or professional environments.

This paper investigates automated scoring of summaries based on six analytic rubrics used in the assessment of college-level writing in English as a second language (EL2). The writing task assesses students' comprehension of complex texts and their ability to produce coherent writing. We build upon the Reading-for-Understanding (RU) cognitive framework (Sabatini et al., 2013) to which we add the Reading-to-Write (RW) element (e.g., Delaney (2008)) in order to analyze automated scoring both in terms of reading comprehension and writing quality.

The contribution of our work is threefold. Firstly, we experiment with regression models to predict six expert-rated analytic scores, and train models that utilize a combination of linguistic features that measure textual cohesion and coherence, as well as reference-based features that compare the summaries against the source texts and expert-compiled reference summaries. Secondly, we carry out a correlation analysis between the text features and analytic scores, discovering patterns that link the RW and RU constructs, including signals of inadequate L2 competence. Lastly, we compile and make available a dataset of expert-rated college-level summaries in EL2.[1]

2 Related Work

Automated scoring of student writing has attracted considerable attention due to the opportunity to analyze cognitive aspects of writing as well as a

[1] http://takelab.fer.hr/el2-summaries

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 181–186,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

need to automate the time-consuming, cognitively demanding, and sometimes insufficiently reliable assessment process, e.g., (Burstein et al., 2013; Rahimi et al., 2015). Much has been done in the area of L1 and L2 essays, e.g., with Coh-Metrix (Crossley and McNamara, 2009; McNamara et al., 2010; Crossley and McNamara, 2011), and some studies have investigated automated scoring also in summaries, e.g., (Madnani et al., 2013). As assignments which demonstrate students' reading/writing skills and their broader academic abilities, summaries have been studied as part of university-level L2 assessment; e.g., integrated task in (Guo et al., 2013).

In such research, holistic scoring mostly supported by well-defined descriptors, e.g., (Rahimi et al., 2015), has predominantly been used to compare against automatically computed features to asses essay quality, e.g., (McNamara et al., 2010), coherence and related concepts such as comprehension in summaries, e.g., (Madnani et al., 2013; Mintz et al., 2014), ease of reading (Burstein et al., 2013) or essay organization. To the best of our knowledge, no research reports have been published on using human raters' multiple analytic scores in such studies.

From a technical perspective, the work most similar to ours is that of Madnani et al. (2013), whose model also uses reference-based features, source-copying features as well as a feature signaling text coherence for automated scoring of summaries. However, they frame the problem as a classification task and predict a single holistic score, whereas we frame the problem as a regression task and predict the scores for six rubrics.

3 Reading-for-Understanding and Reading-to-Write in L2

Summarization can be perceived as a Reading-for-Understanding (RU) task as discussed by Madnani et al. (2013) based on (Sabatini et al., 2013). In other words, summarizing includes lower- and higher-level comprehension processes leading to establishing coherence according to the most plausible intended meaning of the source text (Grosz and Sidner, 1986). Meaning is thus actively constructed by selecting and organizing the main ideas of the text into a coherent whole. When the result of comprehension processes is articulated in writing, there is a need to introduce cohesion devices which signal the rhetorical structure of the text and ensure a smooth flow of sentences. Summary writing is thus also a Reading-to-Write (RW) task (e.g., Delaney (2008)) demonstrating the ability to "convey information" as "a central component of real-world skills" (Foltz, 2016).

While there is a natural overlap between RU and RW (since RW includes and largely depends on RU), the difference between the two constructs is more prominent when summaries are written in L2. For example, a text which is mostly well understood by a non-native speaker may be poorly summarized due to insufficiently developed L2 writing leading to overreliance on bottom-up processing and lack of content integration. The RW manifestation of such problems may be "inability to paraphrase" and plagiarism, poor cohesion (Kirkland and Saunders, 1991), or weak text organization. Conversely, advanced L2 writing may sometimes combine with superficial reading (also seen in native speakers), resulting in factually inaccurate, incomplete, or incoherent summaries.

Analytic scoring based on different rubrics (e.g., accuracy, cohesion) is therefore particularly appropriate when assessing summaries in L2 as it offers more informative feedback (Bernhardt, 2010) and better captures different facets of L2 writing competence than holistic assessment (Weigle, 2002). However, analytic scoring is often exceptionally demanding for raters, especially in the case of longer texts and more than four or five scoring categories (CEFR), which motivates the use of automated assessment.

4 Data Collection

The research encompassed 114 first-year business undergraduates whose competence in English as L2 was predominantly upper intermediate and advanced. Two text-present summary writing tasks (tot. 228 summaries) were administered for two respective articles (ca. 900 words each) taken from *The Economist*, a renowned business magazine. Both times participants were required to read the article and write a summary of about 300 words. Participants were instructed that the summary should clearly present the main ideas to a third person who did not read the article.

In this work, we conceptualize RW as the ability to produce a well-organized writing with well-connected sentences (cohesion), clear paragraphing, topic sentences, and clear links between paragraphs (text organization), while RU is about covering the content of the source text accurately (factually accurate at the local level), completely (all the main ideas covered), relevantly (important ideas included), and coherently (a logical flow of ideas).

Two expert raters assessed the summaries on the 4-point analytic scales (grades 0–3) consisting of the RU content-based (accuracy/Acc, completeness/Cmp, relevance/Rev, and coherence/Chr), and RW text-based rubrics (cohesion/Chs, text organization/Org). The scales were quantified to the extent possible (e.g., by defining the number of cohesion breaks or accuracy errors for each grade). Inter-rater reliability measured by weighted kappas was as follows: accuracy 0.64, completeness 0.76, relevance 0.76, coherence 0.69, text organization 0.76, and cohesion 0.83. Despite adequate reliability, the relatively small number of summaries allowed the raters to discuss and agree all the grades.

Before automated scoring, all the summaries were checked for spelling and basic grammar (e.g., adding "s" to verbs in the present tense of the third person singular), as we were primarily interested in higher-level comprehension processes in the RU/RW construct, and not in grammar or spelling. Also, two reference summaries for each text were written by experts following the same instruction as the one given to students.

5 Automated Scoring

We frame the automated scoring as a multivariate regression task and train separate regression models for each of the six rubrics. Each regression model is trained to predict the expert-assigned score on a 0–3 scale. In using regression instead of classification, we utilize the ordinal nature of the rubric scores, but posit the equidistance of two consecutive scores.

Features. Each of the six regression models is trained on the same set of features. The features can be grouped into reference-based features (BLEU, ROUGE, and "source-copying" features) inspired by (Madnani et al., 2013), and linguistic features derived from Coh-Metrix indices. For preprocessing (sentence segmentation and tokenization), we use the NLTK toolkit of Bird (2006).

- **BLEU** (Papineni et al., 2002) is a precision-based metric originally used for comparing machine-generated translations against reference translations. In our case, BLEU measures the n-gram overlap between the student's summary and the source text. The rationale is that a good summary will refer to the ideas from – and hence likely reuse fragments of – the source text.
- **ROUGE** (Lin and Hovy, 2003), is a recall-oriented metric originally used for evaluating automated summarization systems. Following (Madnani et al., 2013), we use ROUGE to compare the student's summary against the two reference summaries. ROUGE is complementary to BLEU and measures to what extent the student's summary resembles the reference summaries. The intuition is that a good summary should cover all the ideas described in the reference summaries, which will be indicated by a high n-gram overlap between the two. We use five ROUGE variants: ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L, and ROUGE-SU4.
- Complementary to ROUGE, we adopt four source-copying features from (Madnani et al., 2013). **CopiedSumm** and **CopiedText** features are the sum of lengths of all the n-grams of length three or longer that are copied from the original text divided by the length of the summary and the source text, respectively. **MaxCopy** is the length of the longest n-gram copied from the source text. **FirstSent** is the number of source-text sentences that share an n-gram of length at least two with the first sentence of the summary.
- We use **Coh-Metrix** indices (Graesser et al., 2004; McNamara et al., 2014) to measure the cohesion and coherence of the summaries. The Coh-Metrix tool[2] computes a wide range of indices, from which we selected 48: 11 descriptive (DES), 12 referential cohesion (CRF) 8 LSA overlap, 9 connectives (CNC), and 8 situation model (SM) indices.

Models. We use two regression algorithms: an L2-regularized linear regression model (Ridge regression) and a non-linear support vector regression (SVR) machine (Drucker et al., 1997) with an RBF kernel. Both algorithms rely on regularization to alleviate the problem of overfitting and multicollinearity. In addition, we experiment with feature selection based on the F-test for each feature, retaining all, 10, or 5 top-ranked features, yielding six different models. We use the sklearn implementation of the algorithms (Pedregosa et al., 2011).

Setup. We evaluate the models using a nested 10×5 cross-validation, where the inner five folds

[2] http://cohmetrix.com

Model	Acc	Cmp	Rel	Chr	Org	Chs
Baseline	43.5	42.3	46.3	35.8	37.4	36.6
Ridge-all	42.2	51.6	46.8	42.3	39.3	48.1
Ridge-10	**54.1***	**54.1***	46.8	**47.7***	43.2	**55.9***
Ridge-5	**54.1***	50.2	**50.8***	47.3*	**44.5***	53.8*
SVR-all	44.8	47.2	49.4	35.8	39.7	36.6
SVR-10	30.3	37.9	41.3	35.3	28.2*	36.5
SVR-5	29.6*	39.5	35.2	34.4	36.2	37.4

Table 1: Accuracy of automated scoring across the six rubrics for the baseline and the six models using all, 10, and 5 features. Maximum scores for each rubric are shown in bold; "*" indicates statistically significant difference against baseline at p<0.05.

are used to optimize the hyperparameters via grid search. The models' performance is measured in terms of accuracy averaged over the five outer folds, by rounding the predictions to closest integers prior to computing the accuracy scores. All the features are z-scored on the train set, and the same transformation is applied on the test set. As the baseline for each rubric, we use the average expert-assigned score for that rubric. We use a two-tailed t-test to compare against the baseline, after having verified that the normality assumption is met.

Results. Table 1 shows the results. We observe that the performance varies considerably across the models and rubrics. The non-linear SVR models perform rather poorly in comparison to the baseline. On the other hand, ridge regression models with 5 or 10 features (depending on the rubric) outperform the baselines on all the six rubrics (the difference is significant at p<0.05). The improvement is most marked for cohesion and coherence (52% and 33% relative improvement, respectively), while the organization rubric appears to be the most difficult to predict. Feature selection improves the performance of ridge regression, suggesting that feature redundancy persists despite regularization.

6 Correlation Analysis

While the reference-based and linguistic features serve as good predictors for the analytic scores of college-level summaries in EL2, we expected not all the features to be equally important for all the scores. We therefore analyzed the correlations between the rubric scores and features which were ranked among the top five for any of the rubrics, plus the ROUGE-3 feature. Table 2 shows Spearman's rank correlation coefficients.

	Acc	Cmp	Rel	Chr	Org	Chs
BLEU	0.27	−0.38	−0.50	−0.51	−0.49	−0.60
CopiedOrig	0.30	−0.36	−0.48	−0.51	−0.49	−0.61
CopiedSumm	0.32	−0.35	−0.46	−0.52	−0.48	−0.59
MaxCopy	0.29	−0.35	−0.39	−0.40	−0.34	−0.38
CNCAdd		0.36	0.42		0.31	0.39
CNCAll		0.33	0.42	0.30	0.31	0.42
CNCLogic			0.39	0.34	0.40	0.46
CRFAOa			0.31	0.41	0.36	0.44
CRFCWOa			0.31	0.37	0.36	0.42
DESWLlt	0.29	0.28				
DESWLsy	0.28	0.28				
ROUGE-3			−0.30		−0.25	−0.34

Table 2: Correlations between top-ranked features and the six rubrics. Correlations of a magnitude <0.25 are omitted. All shown correlations are statistically significant at p<0.05.

The analysis reveals two correlation patterns between (1) human analytic scoring of RU (e.g., accuracy) and RW (e.g., cohesion), and (2) the computationally derived features (linguistic and reference-based). On the one side, accuracy (local, factual) is the only RU/RW dimension which correlates positively with BLEU and the three source-copying features (CopiedOrig, CopiedSumm, and MaxCopy). Moreover, accuracy and completeness are the only two dimensions positively correlating with word length features, which may also be related to the copying effort (plagiarism) when summarizing a demanding text at lower L2 competence.

On the other side, all the other RU/RW dimensions (completeness, relevance, coherence, organization, and cohesion) correlate negatively with the plagiarism-related indices, as well as with ROUGE-3. Also, positive correlations are found in all the RU/RW dimensions but accuracy with some or all of the indices relating to coherent writing (i.e., CNC connectors and CRF argument and content word overlaps). Furthermore, text organization and cohesion, the RW dimensions in our study, show the same correlation patterns as two key content-based criteria (RU): relevance and coherence.

7 Conclusion

In this paper we considered automated scoring of a college-level summary writing task in English as a second language (EL2), building on the Reading-for-Understanding (RU) cognitive framework to which we added the Reading-to-Write (RW) element. The automated scoring of the summaries was based on six analytic rubrics. A regularized regression model which uses a combination of reference-based and linguistic features outperformed the baseline model across all the six rubrics, yielding accuracies between 44.5% and 55.9% on a 4-point scale.

While this result needs to be improved to be of practical use, we discovered interesting links between RW and RU in L2 and the potential of our system to measure the construct analytically. Local accuracy found in summaries may to a large extent be related to plagiarism as a strategy demonstrating underperforming RW (rather than successful RU) as inadequate L2 and/or less developed academic ability prevent comprehension and paraphrasing. Unlike accuracy, the other dimensions (completeness, relevance, coherence, text organization and cohesion) relate to active meaning construction when building a coherent mental representation (e.g., using connectors to clarify on the links between ideas), either in reading or writing. In RU, this may mean searching for relevant information, and integrating it into a coherent whole, while in RW, coherence is possibly achieved through cohesion and text organization.

Acknowledgments

We thank Svjetlana Kolić-Vehovec for her valuable advice on the construction of the analytic scales used in this study and to Višnja Kabalin-Borenić for her assistance in the assessment of the summaries we analyzed in this work. Our thanks also go to Marin Andrijašević for his technical support.

References

Elizabeth B. Bernhardt. 2010. *Understanding Advanced Second-Language Reading*. Routledge.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Jill Burstein, Joel Tetreault, and Martin Chodorow. 2013. Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, 4(2):34–52.

Scott A. Crossley and Danielle S. McNamara. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18(2):119–135.

Scott A. Crossley and Danielle S. McNamara. 2011. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):170–191.

Yuly Asencion Delaney. 2008. Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, 7(3):140–150.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Peter W. Foltz. 2016. Advances in automated scoring of writing for performance assessment. In *Handbook of Research on Technology Tools for Real-World Skill Development*, pages 659–678. IGI Global.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Liang Guo, Scott A. Crossley, and Danielle S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A somparison study. *Assessing Writing*, 18(3):218–238.

Margaret R. Kirkland and Mary Anne P. Saunders. 1991. Maximizing student performance in summary writing: Managing cognitive load. *Tesol Quarterly*, 25(1):105–121.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated scoring of a summary-writing task designed to measure reading comprehension. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168. Association for Computational Linguistics.

Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Lisa Mintz, Dan Stefanescu, Shi Feng, Sidney D'Mello, and Arthur Graesser. 2014. Automatic assessment of student reading comprehension from short summaries. In *Educational Data Mining 2014*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Zahra Rahimi, Diane J Litman, Elaine Wang, and Richard Correnti. 2015. Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 20–30. Association for Computational Linguistics.

John Sabatini, Tenaha O'Reilly, and Paul Deane. 2013. Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design. *ETS Research Report Series*, 2013(2).

Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge University Press.