Analysing Rhetorical Structure as a Key Feature of Summary Coherence

Jan Šnajder[1] **Tamara Sladoljev-Agejev**[2] **Svjetlana Kolić-Vehovec**[3]
[1] University of Zagreb, Faculty of Electrical Engineering and Computing, TakeLab
[2] University of Zagreb, Faculty of Economics and Business
[3] University of Rijeka, Faculty of Humanities and Social Sciences, Department of Psychology
jan.snajder@fer.hr, tagejev@efzg.hr, skolic@ffri.hr

Abstract

We present a model for automatic scoring of coherence based on comparing the rhetorical structure (RS) of college student summaries in L2 (English) against expert summaries. Coherence is conceptualised as a construct consisting of a rhetorical relation and its arguments. Comparison with expert-assigned scores shows that RS scores correlate with both cohesion and coherence. Furthermore, RS scores improve the accuracy of a regression model for cohesion score prediction.

1 Introduction

Assessment of text quality may benefit from automatic scoring as it is cognitively demanding and often requires much expertise (Rahimi et al., 2017), especially in college-level expository writing. One of the key aspects of text quality is writing coherence (Crossley and McNamara, 2010) which reflects students' ability to connect ideas in their mind and to convey the same message in essays or summaries (Halliday and Hasan, 2014).

Existing approaches to text quality predominantly focus on surface measures for assessment (e.g., number of cohesive devices), which sometimes have little relation either to human judgment, e.g., text length (Mintz et al., 2014), or to text-specific meaning (Rahimi et al., 2017). However, automatic scoring of coherence should ideally provide clear and reliable feedback (Burstein et al., 2013) based on features with cognitive validity, e.g., (Loukina et al., 2015).

One way to meet such requirements is to define coherence as the identification of relations between the text's ideas (Rapp et al., 2007). Such a definition may best be analysed in summaries in which the key ideas of the source text are integrated into a *rhetorical structure* (RS).

In cognitive terms, writing summaries is an exercise in *reading-for-understanding* (RU) (Sabatini et al., 2013) and gist reasoning (Chapman and Mudar, 2013). The result of such processes is the macrostructure of the source text constructed in the reader's mind (Louwerse and Graesser, 2005), which consists of concepts and propositions, their mutual relations (Sanders and Noordman, 2000), and relations with prior knowledge. Coherent summaries should express the intention of the source text (Hobbs, 1993) using linguistic devices (cohesion), which makes summarisation also a *reading-to-write* (RW) task (Delaney, 2008). Moreover, summaries have a distinctive feature for annotation: a largely shared knowledge base, i.e., the source text(s) known both to the writer and to the rater(s), which assists raters in their judgment and helps develop a reliable text-specific scoring tool.

In this paper we present a model for automatic scoring of summaries based on analysing a rhetorical structure of a student's summary compared to that of reference summaries. Our starting point is coherence conceptualized as a construct consisting of three elements: a rhetorical relation and its two arguments. We posit that expository text has a rhetorical structure (RS) consisting of a series of text-specific rhetorical segments, the majority of which will be conveyed in a coherent summary if full text-level comprehension is achieved. The model uses a discourse parser to extract rhetorical structures of summaries, and then compares similarity of these structures. We show that the scores produced by the model correlate with the expert-assigned cohesion and coherence scores as well as with surface indices of cohesion. We also show that the model-produced scores can be used to improve cohesion score prediction.

2 Related Work

Automatic assessment of text quality can include content, language accuracy, sophistication and style

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 46–51
Florence, Italy, August 2, 2019.

as well as sometimes overlapping features such as topic similarity, focus, coherence, cohesion, readability, or text organisation and development, e.g., (Pitler et al., 2010; Yannakoudakis and Briscoe, 2012; Guo et al., 2013; Rahimi et al., 2015; Gao et al., 2018). Coherence is a broad concept assessed by different automatic tools, e.g., (Higgins et al., 2004; Yannakoudakis and Briscoe, 2012; Burstein et al., 2013). Scoring measures may include surface features such as word or text length or the number of pronouns and connectives, e.g., (Yannakoudakis and Briscoe, 2012; MacArthur et al., 2018), which may also be contextualised, e.g., (Pitler et al., 2010). Source overlaps may also be used in scoring such as overlapping n-grams in summaries (Madnani et al., 2013), and semantic similarity (e.g,. LSA) may provide information on relatedness between words, e.g., lexical chaining (Somasundaran et al., 2014), sentences (Foltz et al., 1998; Higgins et al., 2004; Higgins and Burstein, 2007), or larger text sections (Crossley and McNamara, 2010). Both types of features (surface and LSA) are encompassed by Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014), a comprehensive computational tool using a range of measures to grasp cognitive aspects of text analysis. Moreover, inter-sentential coherence can be measured using syntax-based entity grids (Barzilay and Lapata, 2008), for example, to distinguish between high- and low-coherence essays (Burstein et al., 2010), or analysing discourse relations (Pitler and Nenkova, 2008; Skoufaki, 2009).

In order to improve the predictive value of automatic assessment, scoring measures are often combined. For example, Pitler and Nenkova (2008) use entity grids, syntactic features, discourse relations (Prasad et al., 2008), vocabulary, and length features. Yannakoudakis and Briscoe (2012) examine different measures and find that semantic similarity is the best addition to lexical and grammatical features. Somasundaran et al. (2014) combine lexical chains, grammar, word usage, mechanics, and RST discourse relations (Mann and Thompson, 1988) in L1 and L2 texts, while Higgins et al. (2004) use semantic similarity together with discourse structure to measure relatedness to the essay question and between discourse segments. More recently, Sladoljev-Agejev and Šnajder (2017) combine reference-based and linguistic features (e.g., Coh-Metrix, BLEU, ROUGE) to predict coherence and cohesion in college student summaries in L2.

The coherence assessment model presented here relies on summaries as a RU/RW task which consists of detecting and conveying the RS of the source text. Similar to Higgins et al. (2004), we use semantic similarity and rhetorical structure to assess coherence of student summaries against summaries written by experts. While Higgins et al. measured the coherence of functional discourse segments (e.g., thesis, conclusion) via semantic similarity between their respective sentences, in our study coherence is measured via similarity between rhethorical structures. Our intuition relies on the establishment of source macrostructure as a coherence-building exercise during reading. Such an approach appears to be cognitively valid and may ensure meaningful feedback both in terms of comprehension and writing skills development or assessment. Our model is constrained by the source content, so we also compare its performance to cohesion features provided by Coh-Metrix in (Sladoljev-Agejev and Šnajder, 2017) to assess generic RW skills.

3 Summary Scoring Model

The summary scoring model works by comparing the RS of a student summary against the rhetorical structures of one or more reference summaries. The model produces a score that indicates to what extent the two structures overlap.

Discourse parsing. To extract the rhetorical relations and their arguments, we use the PDTB-style parser of Lin et al. (2014), a state-of-the-art, end-to-end parser which labels instances of both implicit and explicit relations as well as their argument spans. The PDTB relation labels are organized in a three-level hierarchy of "sense tags" (Prasad et al., 2008). The parser recognizes the first two levels: relation Category (e.g., *Comparison*) and Type (e.g., *Contrast*). The end-to-end performance of the parser, measured as F1-score under partial argument matching, is 48%. The output of this step is, for each summary S, a set of rhetorical relations $\{r_i\}_i$, where $r_i = (l_i, a_i^1, a_i^2)$ is a relation of class/type label l_i, while a_i^1 and a_i^2 are text segments corresponding to its arguments.

Comparing rhetorical structures. When comparing the similarity of summaries' rhetorical structures, we want the model to assign high scores to pairs of summaries that have many rhetorical relations in common. Of course, we cannot expect the arguments of rhetorical relations to be literally

the same, but, if two relations of the same label are to be considered equivalent, their corresponding arguments should be highly semantically similar. We formalize this intuition by defining the weight w_{ij} between a pair of rhetorical relations $r_i = (l_i, a_i^1, a_i^2)$ and $r_j = (l_j, a_j^1, a_j^2)$ as:

$$w_{ij} = \begin{cases} \frac{1}{2}\big(\mathrm{s}(a_i^1, a_j^1) + \mathrm{s}(a_i^2, a_j^2)\big) & \text{if } l_i = l_j, \\ 0 & \text{otherwise.} \end{cases}$$

where $s(\cdot, \cdot)$ is the semantic similarity between two text segments. In line with much of recent work, we rely on additive compositionality of word embeddings, and compute the semantic similarity as the cosine similarity between averaged word embeddings of the two segments. We use the 300-dimensional skip-gram word embeddings built on the Google-News corpus (Mikolov et al., 2013).[1]

To compute the overlap score between a pair of summaries S_1 and S_2, each consisting of a set of rhetorical relations, we use the maximum bipartite graph matching algorithm (Kuhn, 1955). The graph edges represent pairs of relations (r_i, r_j), $r_i \in S_1$, $r_j \in S_2$, weighted by w_{ij}. Let $n_1 = |S_1|$ and $n_2 = |S_2|$ be the number of rhetorical relations in S_1 and S_2, respectively, and m the maximum matching score between S_1 and S_2. We define the precision (P) and recall (R) of the match as:

$$P = \frac{m - \max(0, n_1 - n_2)}{n_1}$$
$$R = \frac{m - \max(0, n_2 - n_1)}{n_2}$$

The intuition is that precision is maximized if all relations from S_1 are perfectly matched to some relations from S_2, and conversely for recall. The F1-score is the harmonic mean of P and R. Finally, we compute the F1-score of a student's summary S as the mean of pairwise F1-scores between S and both reference summaries.

4 Evaluation

Dataset. For model evaluation, we adopt the dataset of (Sladoljev-Agejev and Šnajder, 2017). The dataset consists of a total of 225 text-present summaries (c. 300 words) of two articles written by 114 first-year business undergraduates in English as L2 (mostly upper intermediate and advanced). Both articles (c. 900 words each) were taken from The Economist, a business magazine. Two expert raters used a 4-point analytic scale (grades 0–3) to assess the summaries in terms of coherence (RU) and cohesion (RW). The scales were quantified by defining the number of coherence and cohesion breaks. Descriptors for each grade included expressions such as "meaningfully related ideas" and "logical sequencing" (for coherence) and "linguistically connected text segments" (for cohesion). Inter-rater reliability (weighted kappas) was 0.69 for coherence and 0.83 for cohesion. The raters discussed and agreed on all the grades although reliability was adequate. As expected, we observe a strong correlation between coherence and cohesion scores (Spearman correlation coefficient of 0.64). All the summaries were checked for spelling and basic grammar. For the two articles from The Economist, two experts with considerable experience with business texts in English wrote 300-word summaries following the same instruction as the students.

Comparison with expert-assigned scores. To assess the validity of the summary scoring model, we measure the correlations of P, R, and F1 scores produced by the model against expert-provided coherence and cohesion scores, considering both Class and Type levels of PDTB relations. Table 1 shows the results. We can make several observations. First, while all the scores correlate positively with both cohesion and coherence, correlation for coherence is consistently lower, possibly due to the role of the raters' prior knowledge, which is unavailable to the model (also note that inter-annotator agreement is lower for coherence than for cohesion). Second, correlation for Type level is consistently lower than for Class level, which can probably be traced to the PDTB parser being less accurate on Type-level relations. Lastly, we note that the highest correlation with both cohesion and coherence is achieved with the F1-score of the Class level model. These results suggest that the proposed summary scoring model is at least partially successful in modeling both cohesion and coherence – and this in spite of the unavoidable errors of the PDTB parser and errors in similarity computations.

Comparison with Coh-Metrix indices. As mentioned in the introduction, a number of studies have used Coh-Metrix cohesion indices as predictors of both cohesion and coherence. In particular, Sladoljev-Agejev and Šnajder (2017) found

[1] https://code.google.com/archive/p/word2vec/

	Class Level			Type Level		
	P@C	R@C	F1@C	P@T	R@T	F1@T
Chs	0.218	0.320	**0.444**	0.207	0.295	0.426
Chr	*0.105*	0.297	**0.381**	*0.071*	0.257	0.337

Table 1: Spearman correlation coefficients between expert-assigned cohesion (Chs) and coherence (Chr) scores and model-produced scores (P, R, and F1) for Class and Type levels of PDTB connectives. The highest correlations for cohesion and correlation are shown in boldface. All correlations except those shown in italics are statistically significant ($p<0.05$).

	Expert scores		Model scores	
Coh-Metrix index	Chs	Chr	F1@C	F1@T
CNCAdd	0.375	0.229	**0.545**	**0.495**
CNCLogic	**0.453**	0.330	0.492	0.409
CNCAll	0.408	0.289	0.477	0.421
CRFAOa	0.430	**0.405**	0.342	0.320
CRFCWOa	0.416	0.364	0.278	0.278

Table 2: Spearman correlation coefficients between Coh-Metrix indices (connectives – CNC, referential cohesion – CRF) and expert-assigned cohesion (Chs) and coherence (Chr) scores as well as model-produced F1 scores at Class level (F1@C) and Type level (F1@T) of PDTB connectives. The highest correlations in each column are shown in boldface. All correlations are statistically significant ($p<0.05$).

modest correlation between expert-assigned coherence/cohesion and indices for connectives (additive connectives – CNCAdd, logical connectives – CNCLogic, and all connectives – CNCAll) and referential cohesion indices (mean of noun/pronoun overlaps between two sentences – CRFAOa, and content word overlap – CRFCWOA). It is therefore interesting to investigate to what extent these surface-level predictors correlate with the scores of our model. Table 2 gives Spearman correlation coefficients between the Coh-Metrix indices and expert-provided scores as well as the Class- and Type-level F1-scores of the model. The Coh-Metrix indices correlate positively with both the expert-assigned scores and the scores of our model. However, while CNCLogic and CRFOAo indices mostly correlate with the expert-assigned cohesion and coherence scores, respectively, the scores of our model mostly correlate with the CNCAdd index.

Supervised scoring. Following Sladoljev-Agejev and Šnajder (2017), we frame the automated scoring as a multivariate regression task and use two regression models, one for cohesion and the other for coherence, each trained to predict the expert-assigned score on a 0–3 scale. We use an L2-regularized linear regression model (ridge regression)[2] and consider three sets of features: (1) five Coh-Metrix CNC and CRF indices ("CM"), (2) the F1-scores of the summary scoring model computed at Class and Type levels ("RS"), and (3) a combination of the two ("CM+RS"). We evaluate the models using a nested 10×5 cross-validation: the models' performance is measured in terms of accuracy averaged over the five outer folds, after rounding the predictions to closest integers and limiting the scores to the 0–3 range. All the features are z-scored on the train set, and the same transformation is applied on the test set. As baselines, we use the rounded average of the expert-assigned scores.

Model / Features	Chs	Chr
Baseline	0.369	0.361
Ridge / CM	0.489	0.409
Ridge / RS	0.476*	**0.419**
Ridge / CM+RS	**0.511***	0.414

Table 3: Accuracy of cohesion (Chs) and coherence (Chr) scores predictions for the baseline and ridge regression models with Coh-Metrix (CM), rhetorical structure (RS), and combined (CM+RS) feature sets. The best results are shown in bold. The "*" indicates a statistically significant difference to baseline ($p<0.05$, Wilcoxon signed-rank test). The differences between regression models with the CM feature set and models with RS and CM+RS feature sets are not statistically significant.

Table 3 shows the results. We can make three main observations. Firstly, cohesion models outperform the corresponding coherence models. Secondly, the only two models for which the differences against the baseline are statistically significant are the two cohesion models that use RS. This suggests that our model does provide useful signals for predicting expert-assigned cohesion scores. In the absence of statistical significance, the results for coherence are inconclusive, though we observe a similar trend.

5 Conclusion

We have described a model for coherence scoring based on a simple definition of coherence in line

[2] We use the implementation of Pedregosa et al. (2011).

with cognitive theories of text comprehension. The model produces scores that correlate with expert-assigned scores and improve the cohesion prediction of a regression model: a model that uses rhetorical structure scores as features yields a statistically significant improvement over the baseline of averaged expert-assigned scores. The proposed model could provide a basis for meaningful feedback in summaries and other similar tasks, and may also be used for measuring gist reasoning in case of a shared knowledge base between the rater and the examinee.

Acknowledgments

We thank Višnja Kabalin-Borenić for her contribution in the assessment of summaries analysed in this work.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 681–684.

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

Sandra Bond Chapman and Raksha Anand Mudar. 2013. Discourse gist: A window into the brains complex cognitive capacity. *Discourse Studies*, 15(5):519–533.

Scott Crossley and Danielle McNamara. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

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

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Yanjun Gao, Patricia M Davies, and Rebecca J Passonneau. 2018. Automated content analysis: A case study of computer science student summaries. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 264–272.

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.

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 comparison study. *Assessing Writing*, 18(3):218–238.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. Routledge.

Derrick Higgins and Jill Burstein. 2007. Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Jerry R Hobbs. 1993. Intention, information, and structure in discourse: A first draft. In *Burning Issues in Discourse, NATO Advanced Research Workshop*, pages 41–66. Citeseer.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Anastassia Loukina, Klaus Zechner, Lei Chen, and Michael Heilman. 2015. Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–19.

M. M. Louwerse and A. C. Graesser. 2005. Macrostructure. *Encyclopedia of Language and Linguistics*.

Charles A MacArthur, Amanda Jennings, and Zoi A Philippakos. 2018. Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, pages 1–22.

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.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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*.

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.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pages 544–554. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn discourse treebank 2.0. In *LREC*. Citeseer.

Zahra Rahimi, Diane Litman, Richard Correnti, Elaine Wang, and Lindsay Clare Matsumura. 2017. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4):694–728.

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.

David N Rapp, Paul van den Broek, Kristen L McMaster, Panayiota Kendeou, and Christine A Espin. 2007. Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific studies of reading*, 11(4):289–312.

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).

Ted JM Sanders and Leo GM Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60.

Sophia Skoufaki. 2009. An exploratory application of rhetorical structure theory to detect coherence errors in L2 English writing: Possible implications for automated writing evaluation software. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 2, June 2009-Special Issue on Computer Assisted Language Learning*, 14(2).

Tamara Sladoljev-Agejev and Jan Šnajder. 2017. Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in l2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 181–186.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, pages 950–961.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43. Association for Computational Linguistics.