Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension

Renu Balyan
Arizona State University Tempe, AZ, USA
renu.balyan@asu.edu

Kathryn S. McCarthy
Arizona State University Tempe, AZ, USA
ksmccar1@asu.edu

Danielle S. McNamara
Arizona State University
Tempe, AZ, USA
dsmcnamara@asu.edu

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2017). Combining machine learning and natural language processing to assess literary text comprehension. In A. Hershkovitz & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (pp. 244-249), Wuhan, China: International Educational Data Mining Society. Published with acknowledgment of federal support.

Author's Note

This research was supported in part by IES Grants R305A150176, R305A130124, and R305A120707, as well as ONR Grants N00014-14-1-0343 and N00014-17-1-2300. Opinions, conclusions, or recommendations do not necessarily reflect the views of the IES or ONR.

Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension

Renu Balyan
Arizona State University
Tempe, AZ, USA
renu.balyan@asu.edu

Kathryn S. McCarthy
Arizona State University
Tempe, AZ, USA
ksmccar1@asu.edu

Danielle S. McNamara
Arizona State University
Tempe, AZ, USA
dsmcnamara@asu.edu

ABSTRACT

This study examined how machine learning and natural language processing (NLP) techniques can be leveraged to assess the interpretive behavior that is required for successful literary text comprehension. We compared the accuracy of seven different machine learning classification algorithms in predicting human ratings of student essays about literary works. Three types of NLP feature sets: unigrams (single content words), elaborative (new) n-grams, and linguistic features were used to classify idea units (paraphrase, text-based inference, interpretive inference). The most accurate classifications emerged using all three NLP features sets in combination, with accuracy ranging from 0.61 to 0.94 (F=0.18 to 0.81). Random Forests, which employs multiple decision trees and a bagging approach, was the most accurate classifier for these data. In contrast, the single classifier, Trees, which tends to "overfit" the data during training, was the least accurate. Ensemble classifiers were generally more accurate than single classifiers. However, Support Vector Machines accuracy was comparable to that of the ensemble classifiers. This is likely due to Support Vector Machines' unique ability to support high dimension feature spaces. The findings suggest that combining the power of NLP and machine learning is an effective means of automating literary text comprehension assessment.

Keywords

Natural language processing; supervised machine learning; classification; interpretation

1. INTRODUCTION

Text comprehension researchers employ a variety of methods to assess how people process and understand the things that they read. The majority of this work has focused on how readers comprehend expository or informational texts (e.g., science textbooks or historical accounts) and simple narratives (e.g., brief plot-based texts). Much less work has been done to investigate the kinds of processes that occur when readers read literary texts, such as the poems, short stories, and novels assigned in English-Language Arts classrooms [1]. More so than in other text domains, literary text comprehension requires the construction of interpretations that go beyond the literal story to speak to a deeper meaning about the world at large [2].

In order to measure interpretation and assess literary comprehension, researchers have relied on collecting students' essays about the text. The essay can then be scored in a variety of ways to address different questions about the comprehension process [3]. Unfortunately, reliably evaluating essays is both time and resource intensive. In other text domains, researchers have begun to develop natural language processing (NLP) tools to automate this scoring [4,5]. With this in mind, our goal was to develop a means of automatically assessing students' essays about literary texts, with particular attention readers' interpretation of a text's potential deeper meaning.

Our purpose was to investigate if NLP and machine learning could be combined and leveraged to accurately predict human ratings of students' essays. We drew upon existing text comprehension research to identify and extract three NLP feature sets that were relevant to literary text comprehension. These feature sets were used to compare seven machine learning classification algorithms in their ability to classify idea units in student essays as literal (paraphrase or text-based inferences) or interpretive.

1.1 Text Comprehension

The field of text comprehension investigates the complex activities involved in how people read, process, and understand text. As people read, they generate a mental representation, or mental model. The quality, structure, and durability of this representation reflect the reader's comprehension of the text [6,7]. A critical aspect of this mental representation is the inclusion of inferences. Inferences connect different parts of the text or connect information from the text to information from prior knowledge. Those who generate more inferences have a more elaborated mental representation [6,7]. Importantly, different types of texts and tasks afford different amounts and types of inferences [8]. For example, readers studying for an upcoming test generate explanatory and predictive inferences, whereas readers reading for fun generate personal association inferences. These different types of inferences suggest readers are engaging in different processes and are constructing different mental representations of the text [9]. Given the importance of inferences in successful text comprehension, a majority of text research is aimed at understanding when and how inferences are constructed [10].

1.2 Literary Comprehension

In the study of literary text comprehension, researchers are interested in interpretive inferences. Interpretive inferences reflect a representation of the author's message or deeper meaning [11]. Take for example, the story of the Tortoise and the Hare. A reader may make text-based inferences to maintain a coherent representation of the events of the text. A reader might generate the inference *The tortoise was able to pass the hare because the hare was sleeping* to explain why the slow tortoise was able to beat the speedy hare. In contrast, a reader might generate an interpretive inference that goes beyond the story world to address the moral or message of the story, such as *It is better for someone to be perseverant than talented.* Research indicates that expert literary readers (e.g., English Department faculty or graduate students) allocate more effort to generating interpretive inferences, whereas

novices, who tend to have less domain-specific reading goals and strategies, tend to merely paraphrase, or restate the plot.

Notably, there is no one "right" interpretation, but rather a multitude of possibilities that may be more or less supportable by the evidence in the text [11,12]. Indeed, some might argue that the moral of the Tortoise and the Hare is not about the tortoise's achievement, but instead reflects a cautionary message about the hare's behavior, such as *People should not be over-confident.* As such, assessing interpretation is more difficult than evaluation of performance in well-defined domains that have a single correct answer. To capture and assess interpretations, researchers have relied on open-ended measures, such as think-aloud protocols, in which readers talk aloud about their processing as they read through the text [13,14,15] and through post-reading essays in which students construct responses to various writing prompts [16]. The transcribed think-aloud data and essays are then parsed into sentences or idea units and scored for the kinds of paraphrases and inferences present. In order to reliably categorize the idea units and essay quality, experts develop and refine a codebook that is then used to train raters. These raters work both independently and collaboratively to reach a satisfactory metric of reliability, such as percent agreement or intra-class correlation.

1.3 Natural Language Processing

More recently, a push has been made to incorporate NLP in text comprehension research [17]. Linguistic features from existing texts are extracted using NLP tools [18]. These tools draw upon corpora of large sets of texts and human ratings to measure aspects of language, such as word overlap, semantic similarity, and cohesion. NLP tools can be used to identify and measure linguistic features that reliably predict human essay ratings [4].

2. DATA & METHODS

2.1 Corpus

The corpus included 346 essays written by college students from two experiments investigating literary interpretation [16,19]. The essays were written about two different short stories from different literary genres (science-fiction, surrealist). In the behavioral experiments, participants received differing reading instructions and writing prompts that biased readers towards paraphrasing or interpretation.

2.2 Human Ratings

Four expert raters scored the set of essays using a previously developed codebook [16]. Essays were parsed into idea units (n = 4,111) and each idea unit was labeled as verbatim, paraphrase, text-based inference, or interpretive inference (Table 1). Given the low amount of verbatim units, verbatim and paraphrase were collapsed into a single paraphrase type.

2.3 Classification Algorithms

Machine learning investigates how machines can automatically learn to make accurate predictions based on past observations. Classification is a form of machine learning that uses a supervised approach. In supervised machine learning, the model learns from a set of data with the class labels already assigned. The model uses this existing classification to make classifications on new data.

Data classification consists of two steps; a learning step (or training phase), and a classification step. In the learning step, a classification algorithm builds a model by "learning from" a training set composed of database tuples, and their associated class labels. A training set may be represented as (X, Y), where X_i is an n-dimensional attribute vector, $X_i=(x_1, x_2,\ldots x_n)$ depicting n measurements made on the tuple from n database attributes, respectively A_1, A_2,..A_n. Each attribute represents a 'feature' of X. Each X_i belongs to a pre-defined class label, represented as Y_i [20]. In the classification step, the trained model is used to predict class labels for a test set of new data set that has not been used during model training. This test data is used to determine the accuracy of a classification algorithm, or *classifier*.

Some of the most commonly used classification algorithms are Naïve Bayes [21], Decision Trees [22], Maximum Entropy [23,24], Neural Networks [25], and Support Vector Machines [26,27]. In addition, researchers also employ ensemble techniques that use more than one of the classifying algorithms. These ensemble algorithms include Bagging [28], Boosting [29], Stacking [30], and Random Forests [31].

2.3.1 Naïve Bayesian

Naïve Bayesian algorithm is based on the Bayes' theorem of posterior probability. It is a probabilistic learning method. It assumes that the effect of an attribute value on a given class is independent of other attributes values [21].

Table 1. Idea unit identification: Definitions and examples
(From McCarthy & Goldman, 2015)

Type	Description	Example from *Harrison Bergeron*	Example from *The Elephant*
Verbatim	Copied directly from the text	*The Handicapper General, came into the studio with a double-barreled ten-gauge shotgun. She fired twice, and the Emperor and the Empress were dead before they hit the floor.*	*The schoolchildren who had witnessed the scene in the zoo soon started neglecting their studies and turned into hooligans. It is reported they drink liquor and break windows. And they no longer believe in elephants.*
Paraphrase	Rewording of the sentences from the text; Summary or combining of multiple sentences from the text	*Then [Harrison] and the ballerina were killed by Diana Moon Glampers, the Handicapper General.*	*After seeing this the students gave up on education became drunks and stopped believing in elephants.*
Text-Based Inference	Reasoning-based on information presented in the story, with some use of prior knowledge; connecting information from two parts of the text	*Diana Moon Glampers killed them because they tried to show their true selves.*	*After being deceived by the fake elephant, the children became poor students, and grew up behaving badly because they were lied to*
Interpretive Inference	Inferences that reflect nonliteral, interpretive interpretations of the text	*It shows what kind of a place the world can turn out to be if we let [the government] get out of control.*	*The theme is that being lied to ends the innocence of the young boys and girls.*

2.3.2 Decision Trees

The Decision Trees learning method approximates discrete-valued target functions. The learned function is represented as a decision tree, which is further represented as a set of if-then rules. Each node in the tree specifies a test of some attribute, and one of the possible values of the attribute represents a branch in the tree. The attribute considered for a node is based on the statistical property, information gain [22].

2.3.3 Maximum Entropy (MaxEnt)

MaxEnt models work on a simple principle, and choose a model that is consistent with all of the given facts. The models are based on what is known, and do not make any assumptions about the unknowns [23,24].

2.3.4 Neural Networks

Neural Networks is a computational approach based on a collection of neural units. It is an attempt to model the information processing capabilities of the human nervous system. These models are self-learning, and use a back-propagation algorithm for updating the weights based on feedback [25,32].

2.3.5 Support Vector Machine (SVM)

SVM constructs a hyperplane that separates the data into classes. SVMs are efficient for high-dimensional feature spaces and are among the best supervised learning algorithms [26,27].

2.3.6 Bagging

Bagging (or Bootstrap Aggregation), is a meta-algorithm that considers multiple classifiers. It creates bootstrap samples of a training set using sampling with replacement. Bagging trains each model in the ensemble using each bootstrap sample, and performs classification based on majority voting from trained classifiers [28].

2.3.7 Boosting

Boosting, a meta-algorithm that incrementally builds an ensemble by iteratively training weak learners or classifiers. While training new models, it emphasizes instances that are misclassified by the previous models. Thus, each model is trained on weighted data from the previous model performance. The final result is the weighted sum of the results of all of the classifiers [29].

2.3.8 Stacking

Stacking (or stacked generalization), combines multiple classifiers generated by different learning algorithms on a single data set. This algorithm works by first generating a set of base-classifiers, and then trains a meta-level classifier to combine the outputs of the base-classifiers [30].

2.3.9 Random Forests

Random Forests (or random decision forest) is designed to overcome the "overfitting" problem of decision trees. Random Forests constructs a multitude of decision trees in the training phase, and uses majority voting for classification [31,33,34].

2.4 Feature Sets

Three NLP feature sets were identified as theoretically relevant to the objective: unigrams, linguistic characteristic scores, and "elaborative" (new) unigrams.

2.4.1 Unigrams

Unigrams are the individual content words present in the idea units. The value of a unigram feature was the frequency of that unigram in the corpus. Some of the most common words appearing in the idea units are *elephant* (>1000), *story* (575), *zoo* (429), *handicap* (361), *government* (323), *believe* (158), and *think* (147).

2.4.2 Linguistic Characteristics

The second set of features considered were the linguistic characteristic scores. Ideas that reflect events from the text are likely to be more concrete, whereas those that are interpretive reflect themes (e.g., freedom, loss of innocence) are more abstract [35]. Thus, both *concreteness* and *imagability* were included as indices. Related to the greater sophistication in interpretive language, we also included *word familiarity* and *age of acquisition*. These linguistics characteristics were derived from merging norms of human ratings from three sources [36,37,38]. Details of merging are provided in appendix 2 of the MRC Psycholinguistic Database User Manual [39]. The characteristics, as defined by McNamara and colleagues [40], appear in Table 2.

Table 2. Descriptions of relevant linguistic characteristics
(From McNamara, Graesser, McCarthy, and Cai, 2014)

Linguistic Characteristic	Description
Concreteness	The degree to which a word is non-abstract
Imagability	How easy it is to construct image of a word in one's mind
Familiarity	How familiar a word is to an adult
Age of Acquisition	The age at which a word first appears in a child's vocabulary

2.4.3 Elaborative n-grams

The third feature set was the frequency of "elaborative" n-grams. These were words (unigrams), two consecutive words (bigrams) or three consecutive words (trigrams) that were new in the sense that they appeared in the idea units, but not in the original story. In addition, frequency of occurrence of a set of cue words or phrases that indicate an interpretive idea unit was included in this feature set.

We used a set of 'R' packages for implementing classification algorithms, and extracting the feature sets. The 'R' packages used for classification include 'RTextTools', 'e1071', 'randomForest', 'nnet', 'MASS', and 'caret'. The packages used for text mining, and extracting n-grams from the idea units and essays were 'tm', 'tau', 'openNLP', 'qdap', and 'quanteda'.

3. EXPERIMENTS & RESULTS

3.1 Feature Selection

The three NLP feature categories (frequency of unigrams, linguistic features of words, and number of "elaborative" n-grams and cue words) were tested in seven experiments.

The total number of unigrams extracted from the idea units was 4,406, resulting in a frequency matrix of 4,111 X 4,406 dimensions. This was more than the number of idea units in the corpus. As a means of reducing the dimensions in the data set, highly correlated unigrams (Pearson $r > .65$) were removed. However, this exercise did not significantly reduce the dimensions. It was noted that many of the unigrams did not appear frequently. Several frequency thresholds were tested to determine a frequency that would reduce dimensions, but not overly affect the accuracy of the model. It was determined that a frequency threshold of 10 was sufficient. Including only those unigrams that appeared in the corpus at least 10 times reduced the feature dimensions from 4,406 to 609.

For the second set of features we considered an initial set of 56 linguistic characteristics. The linguistic features included *concreteness*, *familiarity*, *imagability* and *age of acquisition* scores

for all the words, content words, function words, and all words with or without keywords. These features were extracted using two NLP tools: the Tool for the Automatic Analysis of Lexical Sophistication [41] and the Tool for Automatic Analysis of Text Cohesion [42]. Highly correlated (Pearson r >.85) features were removed, yielding 18 linguistic features for the classification tests.

For the "elaborative" n-grams feature set (unigrams, bigrams, and trigrams present in the idea units, but not the original story and cue words), the bigrams and trigrams were found to be highly correlated (Pearson r > 0.85). Consequently, only trigrams were included. In total, three features were used in the elaborative n-gram feature set for classification.

This final feature set was used to classify each idea unit as paraphrase, text-based inference, or interpretive inference using ML classification algorithms. Similar approaches have been used to classify other kinds of texts [43].

3.2 Idea Unit Classification

After experimenting with a large number of classification algorithms, we selected four machine learning classification algorithms (Trees, Support Vector Machine [SVM], Neural Networks, Maximum Entropy [MaxEnt]), as well as three ensemble approaches (Bagging, Boosting, Random Forests) to classify the idea units. Multiclass classification algorithms and 10-fold cross-validation were used in seven experiments to test the feature sets (609 unigrams, 18 linguistic features, and 3 elaborative n-grams) individually and in combination. Summary of classification accuracy for all the algorithms is presented in Table 3.

The bold entries in Table 3 indicate the maximum accuracy for each of the features. Random Forests achieved the highest accuracy for all experiments except when using elaborative n-grams as features. The Boosting algorithm classifier achieved the maximum accuracy in this case.

The italicized entries in Table 3 indicate the maximum accuracy achieved by a classification algorithm. Generally, the classification algorithms achieved high accuracy when a combination of all features was used. The accuracy for the algorithms varied between 0.77 and 0.94 when considering a combination of all the features, except for the Trees algorithm where the accuracy was quite low, 0.61. In fact, the accuracy for the Trees algorithm was low in all cases irrespective of the features considered.

F-scores for the three types of idea units produced by participants (interpretive, paraphrase, text-based) are summarized in Tables 4 and 5 for single classifiers and ensemble of classifiers, respectively. The bold numbers indicate the highest F-score for each type of idea unit. For the single classifiers, SVM achieved the highest F-score for paraphrases (F = 0.81) and for interpretive inferences (F = 0.73). MaxEnt obtained the highest F-score for single classifiers for text-based inferences (F = 0.42). For ensemble classifiers, Random Forests again performed the best, with the highest F-scores for paraphrases (F = 0.80) and interpretive inferences (F = 0.70). The Bagging algorithm achieved the highest F-score (0.30) for text-based inferences in ensemble category. The F-scores for identifying text-based inferences were relatively low, suggesting a machine learning approach may be better suited for identifying paraphrases and interpretations. The NAs in Table 4 indicate that the algorithm did not classify any idea unit as text-based.

Table 3. Accuracy for different classification algorithms with different feature combinations

[1]Unigrams (n=609); [2]Linguistic Features (n=18); [3]Elaborative n-grams (n=3; unigrams, trigrams, cue words)

	Classification Algorithm						
Feature	**SVM**	**Trees**	**MaxEnt**	**NeuralNets**	**Boosting**	**Bagging**	**Random Forests**
UNI[1]	0.75	0.58	0.81	*0.77*	0.73	0.75	**0.86**
LIN[2]	0.80	0.56	0.55	0.58	0.77	0.92	**0.94**
ENC[3]	0.64	0.60	0.58	0.62	***0.79***	0.63	0.61
UNI + LIN	0.77	0.58	*0.83*	0.76	0.74	0.92	***0.95***
UNI + ENC	0.78	*0.61*	0.80	*0.77*	0.77	0.82	**0.88**
LIN + ENC	*0.92*	0.59	0.62	0.63	*0.79*	*0.93*	**0.94**
UNI + LIN+ ENC	0.81	*0.61*	0.82	*0.77*	*0.79*	*0.93*	**0.94**

Table 4. F-Scores for Single classifiers

[1]Unigrams (n=609); [2]Linguistic Features (n=18); [3]Elaborative n-grams (n=3; unigrams, trigrams, cue words); [4]Interpretive; [5]Paraphrase; [6]Text-based Inference

	SVM			Trees			MaxEnt			NeuralNets		
Feature	**Inter[4]**	**Para[5]**	**TB[6]**	**Inter**	**Para**	**TB**	**Inter**	**Para**	**TB**	**Inter**	**Para**	**TB**
UNI[1]	0.71	0.80	0.28	0.44	0.71	NA	0.65	0.76	0.36	0.63	0.76	0.13
LIN[2]	0.45	0.73	0.13	0.27	0.70	NA	0.52	0.66	0.30	0.46	0.73	NA
ENC[3]	0.46	0.73	0.03	0.52	0.73	NA	0.50	0.72	NA	0.57	0.74	NA
UNI + LIN	0.70	**0.81**	0.35	0.49	0.72	NA	0.66	0.77	0.41	0.64	0.79	0.08
UNI + ENC	**0.73**	**0.81**	0.34	0.55	0.74	NA	0.69	0.78	0.38	0.62	0.73	0.18
LIN + ENC	0.48	0.73	0.11	0.50	0.73	NA	0.58	0.74	0.25	0.61	0.77	NA
UNI+LIN+ENC	0.72	**0.81**	0.36	0.55	0.74	0.30	0.70	0.79	**0.42**	0.63	0.79	0.06

Table 5. F-Scores for Ensemble classifiers

[1]Unigrams (n=609); [2]Linguistic Features (n=18); [3]Elaborative n-grams (n=3; unigrams, trigrams, cue words); [4]Interpretive; [5]Paraphrase; [6]Text-based Inference

	Boosting			Bagging			Random Forests		
Feature	**Inter[4]**	**Para[5]**	**TB[6]**	**Inter**	**Para**	**TB**	**Inter**	**Para**	**TB**
UNI[1]	0.65	0.77	0.06	0.65	0.76	0.17	0.68	0.79	0.27
LIN[2]	0.49	0.70	0.09	0.51	0.72	0.26	0.51	0.74	0.21
ENC[3]	0.52	0.73	0.06	0.51	0.73	0.18	0.53	0.74	0.02
UNI + LIN	0.57	0.73	0.12	0.61	0.76	0.27	0.67	0.78	0.23
UNI + ENC	0.62	0.76	0.07	0.66	0.77	0.27	**0.70**	**0.80**	0.28
LIN + ENC	0.55	0.73	0.23	0.57	0.75	0.25	0.58	0.77	0.21
UNI +LIN + ENC	0.61	0.76	0.18	0.63	0.79	**0.30**	0.67	0.79	0.23

4. CONCLUSIONS

This study demonstrates that a classification approach using unigrams, linguistic features, and "elaborative" n-grams can be used to accurately predict human ratings of idea unit classification for essays about literary works.

This study indicated that ensemble classification algorithms were, generally, more accurate than single classifiers. Random Forests, which is an ensemble of decision trees and uses a bagging approach, was the most accurate classifier and had the highest F-scores for most types of idea units. In contrast, the single classifier Trees showed relatively low accuracy. This finding is consistent with previous work that suggests Trees "overfits" to training data and, as a result, performs poorly on test data [44].

Interestingly, performance from the single classifier SVM was comparable to the ensemble classifiers. This classifier may have been highly accurate due to the fact that our data had a large amount of features under consideration. SVM is designed to support high-dimension spaces and data that may not be linearly separable.

This study provides a model for how machine learning and NLP can be used to assess literary text comprehension. In addition to being economical for researchers recruiting large samples and collecting large amounts of essay data, the approach can also be implemented in other automated writing evaluators (AWEs) to provide domain-specific assessment and feedback.

The presence of interpretive inferences suggests that a reader has successfully moved beyond the literal to engage in domain-appropriate interpretations. However, interpretive inferences are not necessarily indicative of higher quality literary text comprehension. Literary comprehension requires not only generating interpretations, but also justifying those interpretations with evidence from the text as well as appeals to cultural and literary norms [1,45]. Hence, good essays are likely to have a relatively even distribution of the various types of ideas (e.g., both inferences and interpretations). Our future plans include assessing the essays holistically and develop algorithms to predict those scores. Our ultimate objective is to better understand the relations between idea unit types and essay quality as well as to further the development of automated assessment of literary comprehension.

5. ACKNOWLEDGMENTS

This research was supported in part by IES Grants R305A150176, R305A130124, and R305A120707, as well as ONR Grants N00014-14-1-0343 and N00014-17-1-2300. Opinions, conclusions, or recommendations do not necessarily reflect the views of the IES or ONR.

6. REFERENCES

[1] McCarthy, K. S. 2015. Reading beyond the lines: A critical review of cognitive approaches to literary interpretation and comprehension. *Scientific Study of Literature* 5, 1(Jan. 2015), 99-128.

[2] Goldman, S. R., McCarthy, K. S., and Burkett, C. 2015. Interpretive inferences in literature. In, *Inferences during reading*, E. O'Brien, A. Cook, and R. Lorch, Eds. Cambridge University Press, New York, NY. 386-415.

[3] McCarthy, K. S., Kopp, K. J., Allen, L. K., and McNamara, D. S. under review. Methods of studying text: Memory, comprehension, and learning. In *Handbook of Research Methods in Human Memory*, H. Otani and B. Schwartz, Eds. Routledge.

[4] Crossley, S., Kyle, K., Davenport, J., and McNamara, D. S. 2016. Automatic assessment of constructed response data in a chemistry tutor. In *Proceedings of the 9th International Conference on Educational Data Mining*, T. Barnes, M. Chi, & M. Feng, Eds. (Raleigh, NC, June 29 – July 2, 2016). EDM'16, International Educational Data Mining Society, 336-340.

[5] Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., Griffin, T. D., and Britt, M. A. 2017. Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education* (2017), 1-33.

[6] Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychol. Rev.* 95 (Apr. 1998), 163-182.

[7] Kintsch, W. 1998. *Comprehension: A paradigm for cognition*. Cambridge University Press, Cambridge, England.

[8] Van den Broek, P., Young, M., Tzeng, T., and Linderholm, T. 1999. The Landscape Model of reading: Inferences and the online construction of memory representation. In *The construction of mental representations during reading*, H. van Oostendorp and S. R. Goldman, Eds. Psychology Press, 1999. 71–98.

[9] Van den Broek P., Lorch, R.F., Linderholm, T., and Gustafson, M. 2001. The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition* 29, 8 (Dec. 2001), 1081-1087.

[10] McNamara, D. S. and Magliano, J. P. 2009. Towards a comprehensive model of comprehension. In B. Ross (Ed.), *Psychol Learn. Motiv*. 51 (Dec. 2009), Elsevier Science. New York, NY, 297-384.

[11] Langer, J. A. 2010. *Envisioning Literature: Literary understanding and literature instruction, 2nd edition*. Teachers College Press, New York, NY.

[12] Levine, S. and Horton, W. S. 2013. Using affective appraisal to help readers construct literary interpretations. *Scientific Study of Literature* 3, 1 (Jan. 2013), 105-136.

[13] Burkett, C. and Goldman, S. R. 2016. "Getting the Point" of Literature: Relations Between Processing and Interpretation. *Discourse Processes* 53, 5-6 (Jul. 2016), 457-487.

[14] Graves, B. and Frederiksen, C. H. 1991. Literary expertise in the description of fictional narrative. *Poetics* 20, 1(Feb. 1991), 1-26.

[15] Zeitz, C. M. 1994. Expert-novice differences in memory, abstraction, and reasoning in the domain of literature. *Cognition and Instruction* 12, 4(Dec. 1994), 277-312.

[16] McCarthy, K. S. and Goldman, S. R. 2015. Comprehension of short stories: Effects of task instructions on literary interpretation. *Discourse Processes* 52, 7 (Oct. 2015), 585-608.

[17] Crossley, S. A., Allen, L. K., and McNamara, D. S. 2014. Analyzing discourse processing using a simple natural language processing tool (SiNLP). *Discourse Processes* 51, 5-6 (Jul. 2014), 511-534.

[18] Jurafsky, D. and Martin, J. H. 2009. *Speech and Language Processing, 2nd edition*. Prentice-Hall, NJ.

[19] McCarthy, K.S. and Goldman, S. R. in prep. Effects of Genre Familiarity on Interpretive Behavior.

[20] Han, J., Kamber, M., and Pei, J. 2012. *Data Mining Concepts and Techniques, 3rd edition*. Elsevier.

[21] McCallum, A. and Nigam, K. 1998. *A comparison of event models for naive Bayes text classification*. In AAAI-98 Workshop on Learning for Text Categorization, Tech. rep. WS-98-05, AAAI Press.

[22] Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill, New York.

[23] Rosenfeld, R. 1994. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Doctoral thesis, Carnegie Mellon University.

[24] Ratnaparkhi, A. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Doctoral thesis, University of Pennsylvania.

[25] Zhang, G. P. 2000. Neural Networks for Classification: A Survey. *Trans. Sys. Man Cyber* Part C 30, 4 (November 2000), 451-462.

[26] Joachims, T. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of 10th European Conference on Machine Learning* (April 21-23). ECML'98. Springer-Verlag London, UK, 137-142.

[27] Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management* (Bethesda, Maryland, USA, November 02 - 07, 1998). CIKM'98. ACM, New York, NY, 148-155.

[28] Breiman, L. 1996. Bagging predictors. *Machine Learning* 24, 2 (Aug. 1996), 123-140.

[29] Krogh, A. and Vedelsby, J. 1994. Neural network ensembles, cross validation, and active learning. In *Proceedings of 7th International Conference on Neural Information Processing Systems* (Denver, Colorado). NIPS'94. MIT Press Cambridge, MA, USA, 231-238.

[30] Wolpert, D. 1992. Stacked generalization. *Neural Networks* 5, 2, 241–260.

[31] Schapire, R. E. and Singer, Y. 1999. BoosTexter: A boosting-based system for text categorization. *Machine Learning* 39, 2-3, 135-168.

[32] Rojas, R. 1996. *Neural Networks - A Systematic Introduction*. Springer-Verlag, Berlin.

[33] Schölkopf, B. and Smola, A. J. 2002. *Learning with Kernels*. MIT Press, Cambridge, MA.

[34] Ho, T. K. 1995. Random Decision Forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Montreal, QC, August 14-15, 1995). ICDAR'95, IEEE Computer Society Washington, DC, USA, 278–282.

[35] Rabinowitz, P. 1987. *Before reading: Narrative conventions and the politics of interpretation*. Ohio State University Press, Columbus, Ohio.

[36] Paivio, A., Yuille, J. C., and Madigan, S. A. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psycho*. 76, 1p2 (Jan. 1968), 1-25.

[37] Toglia, M. P. and Battig, W. F. 1978. *Handbook of semantic word norms*. Lawrence Erlbaum.

[38] Gilhooly, K. J. and Logie, R. H. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behav. Res. Methods & Instrum*. 12, 4, 395-427.

[39] Coltheart, M. 1981. The MRC Psycholinguistic Database. Q. *J. Exp. Psychol*. A 33, 4, 497–505.

[40] McNamara, D. S., Graesser, A. C., McCarthy, P., and Cai, Z. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, Cambridge.

[41] Kyle, K. and Crossley, S. A. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49, 4 (Dec. 2015), 757-786.

[42] Crossley, S. A., Kyle, K., and McNamara, D. S. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behav. Res. Methods*. 48, 4 (Dec. 2016), 1227-1237.

[43] Jarvis, S. and Crossley, S. (Eds.). 2012. *Approaching language transfer through text classification: Explorations in the detection-based approach*. Bristol, UK: Multilingual Matters.

[44] Chen, C., Liaw, A., and Breiman, L. 2004. *Using random forest to learn imbalanced data*. University of California, Berkeley, 110.

[45] Sosa, T., Hall, A. H., Goldman, S. R., and Lee, C. D. 2016. Developing symbolic interpretation through literary argumentation. *J. Learn. Sci.* 25, 1 (Dec. 2015), 93-132.