Exploring the Intersection of Short Answer Assessment, Authorship Attribution, and Plagiarism Detection

Björn Rudzewitz
University of Tübingen
Nauklerstrasse 35
72074 Tübingen, Germany
brzdwtz@sfs.uni-tuebingen.de

Abstract

In spite of methodological and conceptual parallels, the computational linguistic applications short answer scoring (Burrows et al., 2015), authorship attribution (Stamatatos, 2009), and plagiarism detection (Zesch and Gurevych, 2012) have not been linked in practice. This work explores the practical usefulness of the combination of features from each of these fields for two tasks: short answer assessment, and plagiarism detection. The experiments show that incorporating features from the other domain yields significant improvements. A feature analysis reveals that robust lexical and semantic features are most informative for these tasks.

1 Introduction

Despite different ultimate goals, Short Answer Assessment, Plagiarism Detection, and Authorship Attribution are three domains of Computational Linguistics that share a range of methodology. However, these parallel have not been compared across domains. This work explores the intersection of these areas in a practical context.

In the domain of authorship attribution, a set of texts and potential authors is given, and the goal is to "distinguish between texts written by different authors" (Stamatatos, 2009, page 1). In the domain of short answer assessment, tools are designed to assess the meaning of a short answer by comparing it to a reference answer (Burrows et al., 2015; Ziai et al., 2012), and thereby to its semantic appropriateness. In the domain of Plagiarism Detection, two main goals can be pursued (Clough and Stevenson, 2011): in extrinsic plagiarism detection, a source and potentially plagiarized texts are compared as a whole unit with methods from the domain of authorship attribution (Grieve, 2007). The goal of intrinsic plagiarism detection is to detect stylistic changes within one document (Zu Eissen and Stein, 2006).

All three areas use textual similarity features on various levels of linguistic abstraction for nominal classifiers, but the distribution of features over three related dimensions differs (Zesch and Gurevych, 2012): style, content, and structure. While (learner language) short answer assessment systems put emphasis on content and ignore stylistic aspects, authorship attribution focuses on stylistic features. Plagiarism detection systems use both content, structural, and stylistic similarity features to classify texts as plagiarizing other documents or not. The main task for short answer assessment and plagiarism detection is to evaluate the existence and quality of paraphrases of a source text. This work explores the effect of features used in the field of authorship attribution and plagiarism detection features for short answer assessment, as well as the effect of short answer assessment features for plagiarism detection.

2 Data

For the experiments in the domain of short answer assessment, the Corpus of Reading comprehension Exercises in German (Ott et al., 2012) was used. For the experiments in the domain of plagiarism detection, the Wikipedia Reuse Corpus (Clough and Stevenson, 2011) was selected for the experiments.

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 235–241,
San Diego, California, June 16, 2016.

These resources were chosen since they are standard shared evaluation resources in these domains (Burrows et al., 2015; Zesch and Gurevych, 2012).

2.1 CREG

CREG-1032 is a short answer learner corpus containing student and reference answers to questions about reading comprehension texts. The longitudinal data was collected at two German programs in the United States at the Ohio State University (OSU) and the Kansas University (KU). The corpus exhibits a high variability of surface forms and semantic content in the student answers due to a variety of proficiency levels represented. Each student answer was annotated by two independent annotators with a binary diagnosis indicating the semantic correctness of the answer, independent of surface variations such as spelling mistakes or agreement errors. The corpus is balanced with respect to this diagnosis. Table 1 shows the distribution of student answers, target answers, and questions, as described in (Meurers et al., 2011b), who also showed that the OSU answers are significantly longer (average token length of 9.7 for KU versus 15.0 for OSU).

2.2 Wikipedia Reuse Corpus

The Wikipedia Reuse Corpus (WRC, (Clough and Stevenson, 2011)) represents different types of text reuse imitating different plagiarism types: copy and paste, light and heavy revision, and non-plagiarism. The plagiarism samples vary in the amount of revision and paraphrasing performed by participants. Table 1 shows the corpus' data distribution. The texts were not exclusively written by English native speakers and show similar surface/semantic variation as the CREG answers. With an average of 208 tokens in length, the answers are nearly 20 times as long as the answers in the CREG corpus, but referred to as "short answers" (Clough and Stevenson, 2011, page 1). Since Zesch and Gurevych (2012) showed empirical deficits in the text reuse conditions, all plagiarism labels were collapsed into a single category, rendering the task a binary classification, parallel to the CREG binary diagnoses. In this setting, the data is unbalanced: the majority class is the plagiarism class with 57 instances, whereas there are only 38 non-plagiarized documents.

	CREG-1032-KU	CREG-1032-OSU	WRC
# student answers	610	422	95
# target answers	136	87	5
# questions	117	60	5

Table 1: Data distribution in the CREG-1032 and Wikipedia Reuse Corpus data set.

3 Baseline Short Answer Assessment System

The UIMA-based CoMiC system (Meurers et al., 2011a; Meurers et al., 2011b) served as a framework for the experiments. It is an alignment-based short answer assessment system which aligns student to reference answers on different levels of linguistic abstraction in order to classify learner answers as (in)correct based on the quantity of different alignment types. CoMiC proved to be highly effective for both German and English (Burrows et al., 2015). The CoMiC system follows a three-stage pipeline architecture (Bailey and Meurers, 2008; Meurers et al., 2011a): alignment, annotation, diagnosis.
First, the system enriches the raw answer texts with linguistic annotation. Table 2 from (Meurers et al., 2011b) shows the different annotation tasks together with the respective tools.

Task	NLP Tool
Sentence Detection	OpenNLP (Baldridge, 2005)
Tokenization	OpenNLP (Baldridge, 2005)
Lemmatization	TreeTagger (Schmid, 1994)
Spell Checking	Edit distance (Levenshtein, 1966) igerman98 word list
POS Tagging	TreeTagger (Schmid, 1994)
NP Chunking	OpenNLP (Baldridge, 2005)
Lexical Relations	GermaNet (Hamp and Feldweg, 1997)
Similarity Score	PMI-IR (Turney, 2001)
Dependency Parsing	MaltParser (Nivre et al., 2007)

Table 2: NLP tools used in the CoMiC system.

In the second step, a globally optimal alignment configuration is selected by the Traditional Marriage Algorithm (Gale and Shapley, 1962). The system aligns tokens, NP chunks, and dependency triples. Tokens are aligned when they match on the surface, lowercased surface, synonym, semantic type, or lemma level. Only new elements (not verbatim given in the corresponding question) are aligned.
In the final step, a range of features (Table 3, (Meurers et al., 2011b)) are extracted and fed to a machine learning component. In contrast to the original

CoMiC system, predictions are made with WEKA's (Hall et al., 2009) memory based learner instead of the TiMBL memory based learner (Daelemans et al., 2007). The features denote directionalized quantities of alignments on different linguistic levels ('pct' = 'percentage of').

Feature	Description
1. Keyword Overlap	pct keywords aligned
2. Target Token Overlap	pct aligned target tokens
3. Learner Token Overlap	pct aligned student tokens
4. Target Chunk Overlap	pct aligned target chunks
5. Learner Chunk Overlap	pct aligned student chunks
6. Target Triple Overlap	pct aligned target dependency triples
7. Learner Triple Overlap	pct aligned student dependency triples
8. Token Match	pct token-identical token alignments
9. Similarity Match	pct similarity-resolved token alignments
10. Semtype Match	pct type-resolved token alignments
11. Lemma Match	pct lemma-resolved token alignments
12. Synonym Match	pct synonym-resolved token alignments
13. Variety of Match (0-5)	sum of features 8-12
14. Target Answer ID	target answer id
15. Student Answer ID	student answer id

Table 3: CoMiC system features.

4 Extensions of the Baseline System

Stamatatos (2009) provides an extensive overview about approaches and stylometric features used in computerized authorship attribution. The features are divided into four subclasses. Table 4 based on (Stamatatos, 2009, page 3) lists all the features used, as well as their corresponding category (lexical/character/syntactic/semantic) and information about whether they are applied to one or two documents. If they are applicable to one document, then there exists a feature both for the student and for the target side in order to model the relation in this specific dimension of similarity, reflected in the prefix 'Student' or 'Target' in the feature names. Features applied to two documents are computed via cosine similarity between a vector for each answer holding the frequencies of the elements under consideration. The feature *all features interpolated* is a special overlap feature, for which first all frequencies of all feature extractors were added to one vector before the cosine similarity was applied (see Figure 1). The first *m* entries in the vector are lexical features, followed by *n* character features, etc.

The *SpellCorr* feature measures the token overlap between two texts using spelling corrected and surface forms. For each token, the system checks whether the token or its lemma appears in a word list. If not, the system seaches the closest Levensthein match cosidering both the other document and the word list. All *.arff* feature files were generated with the same givenness constraints as the CoMiC baseline features and exported from there to WEKA.

Feature	Description	# Docs
lexical		
AvgWordLength	Average word length	1
TTR	Type-Token Ratio	1
WordUniFreq	Word Unigram frequency similarity	2
WordBiFreq	Word Bigram frequency similarity	2
WordTriFreq	Word Trigram frequency similarity	2
SpellCorr	Spell Corrected Unigram Matches	1
character		
CharFreq	Character frequency similarity	2
UpperCharFreq	Uppercase character frequency similarity	2
LowerCharFreq	Lowercase character frequency similarity	2
DigitCharFreq	Digit character frequency similarity	2
LetterProportion	Proportion of letters (A-Za-z) in answer	1
UpperProportion	Proportion of uppercase letters in answer	1
LowerProportion	Proportion of lowercase letters in answer	1
CharBigramFreq	Character bigram frequency similarity	2
CharTrigramFreq	Character trigram frequency similarity	2
CharFourgramFreq	Character fourgram frequency similarity	2
CharFivegramFreq	Character fivegram frequency similarity	2
syntactic		
POS	Part of Speech tag frequency similarity	2
Chunk	Chunk tag frequency similarity	2
NPChunk	Noun phrase chunk frequency similarity	2
PosBigram	POS tag bigram frequency similarity	2
PosTrigram	POS tag trigram frequency similarity	2
PosFourgram	POS tag fourgram frequency similarity	2
PosFivegram	POS tag fivegram frequency similarity	2
semantic		
Synonym	Proportion of synonym-overlapping tokens	1
DepTriple	Proportion of dependency triple overlaps	1
combination		
all features interpolated	all features combined	2

Table 4: Authorship attribution features implemented in CoMiC.

5 Experimental Testing

The following orthogonal hypotheses were tested:

1. The accuracy for the learner language short answer assessment task increases when features from the domain of authorship attribution are added.

2. The accuracy for the plagiarism classification task increases when features from the short answer assessment system are added.

5.1 Method

The WEKA lazy iBk memory based learner with k=5-nearest neighbor search was run in a 10-fold cross validation setting. Following Dietterich (1998), the McNemar's test with $\alpha = 0.1$ is used in

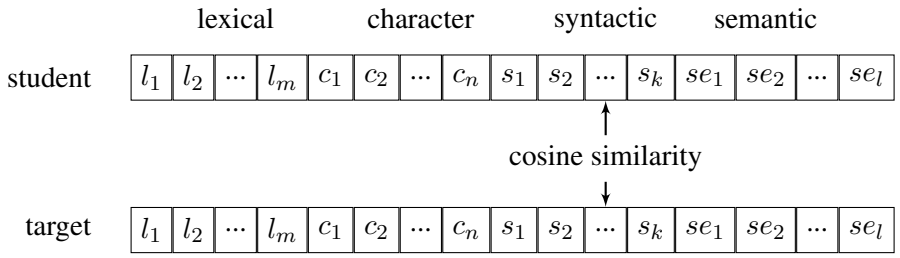

Figure 1: Interpolated textual similarity.

R to test whether an improvement over the baseline is statistically significant.

5.2 Results

Table 5 shows accuracies for the prediction of the semantic equivalence of learner answers and the prediction of plagiarism. For short answer assessment, the CoMiC features yielding an accuracy of 84.5% (KU) and 87.1% (OSU) are used as a baseline. For plagiarism detection, the set of all style features (Table 4, 84.2%) was used as the baseline. Table 5 shows that using only the baseline features from the other domain already yield significant improvements (92.6% for the WRC, 86.9% for CREG-1032-KU) over the baseline of in-domain features. Also the combination of both baseline feature sets yields improvements over the respective baseline. Even though the interpolated similarity feature on its own resulted in a surprisingly high accuracy for CREG-1032-OSU (87.9%), it only works in combination for the WRC corpus, resulting in the highest accuracy of all experiments (95.8%). Lexical features alone result in accuracies comparable to the baseline accuracies for both tasks. The character based features alone work better for short answer assessment, with even better results when combined with the baseline features. Semantic features have a higher impact for plagiarism detection, although for the CREG-1032-OSU data set, these features alone yield nearly the baseline accuracy.

Feature Analysis. The information gain of features was computed in WEKA with the InfoGainAttributeEval filter with default parameters. Table 6 shows the ten most informative features for each data set. The most informative features are mostly lexical or character-based and thus content-modeling features, where the most informative feature indicates the proportion of matched tokens when spelling-corrected versions are used. This is not surprising given the high surface variability in the corpora, and the design choices of the corpus creation to ignore form errors and focus on semantics.

Features	**Data**		
	KU	OSU	WRC
baselines			
CoMiC	84.5	87.1	92.6*
all style features	86.9*	86.0	84.2
baselines + new features			
CoMiC + all style features	85.6	87.7	90.5*
all features interpolated	78.0	87.9	62.1
CoMiC + all features interpolated	84.3	87.2	95.8*
lexical features	84.5	86.3	90.5*
CoMiC + lexical features	84.4	88.2	88.4
character features	83.3	86.3	82.2
CoMiC + character features	85.7	87.7	83.2
syntactic features	67.4	69.0	80.0
CoMiC + syntactic features	84.3	85.1	87.4
semantic features	82.1	85.0	90.6*
CoMiC + semantic features	83.8	87.0	91.6*

Table 5: Results for the binary classification tasks. * denotes a significant improvement ($\alpha = 0.1$).

6 Discussion and Related Work

Grieve (2007) provided an extensive comparison of quantitative authorship attribution methods for extrinsic plagiarism detection. The observation that word and character-based metrics are most successful for extrinsic plagiarism detection can be confirmed by the present study. Clough and Stevenson (2011) tested two methods for classifying the texts in their Wikipedia Reuse Corpus: n-gram overlap and longest common subsequence. They report on an accuracy of 80% for predicting all four labels, and an accuracy of 94.7% for the binary classification. The present work outperformed the already very accu-

Rank	CREG-1032	WRC
1	TargetSpellCorr	StudentSpellCorr
2	CharBigramFreq	Token Match
3	CharTrigramFreq	CharTrigramFreq
4	CharFourgramFreq	CharFourgramFreq
5	WordUniFreq	CharFivegramFreq
6	all features interpolated	TargetSpellCorr
7	Target synonym overlap	CharBigramFreq
8	CharFivegramFreq	StudentSynonym
9	StudentSpellCorr	WordUniFreq
10	TargetSynonym	TargetSynonym

Table 6: Ten most informative features for the CREG-1032 and WRC data set.

rate system by Clough and Stevenson (2011) by almost one percent point with an accuracy of 95.8%. Zesch and Gurevych (2012) used a variety of content, structural, and stylistic features for the plagiarism classification task on the Wikipedia Reuse Corpus. They report an accuracy of 96.8% for the task of binary plagiarism classification.

Meurers et al. (2011b) reported an accuracy of 84.6% for both the CREG-1032-KU and the CREG-1032-OSU data set with an early version of the CoMiC-DE system. Hahn and Meurers (2012) report an accuracy of 86.3% for the CREG corpus as a result of using the CoSeC system, which uses abstract semantic representations. Horbach et al. (2013) re-implemented the CoMiC system and tested the effect of considering the text instead of pre-defined target answers. In the best case, they reached an accuracy of 84.4% on the CREG corpus. Pado and Kiefer (2015) classified answers in the CREG corpus according to their similarity to a target answer. All answers above a threshold were classified as correct, resulting in an accuracy of 83.7% for CREG-1032. Ziai and Meurers (2014) made use of human-annotated information structural annotations for the CREG-1032-OSU data set. They obtained an accuracy of 90.3% for the CREG-1032-OSU data set for the CoMiC system. Rudzewitz (2015) augmented the CoMiC system with alignment weighting features measuring the importance of aligned elements with respect to the concrete task and general linguistic properties of aligned elements. This work reported an accuracy of 90.0% for the CREG-1032-OSU corpus. The difference of 1.2% to the present work warrants a combination of both approaches in future work.

7 Conclusions and Future Work

This article represents a pioneer work for linking the three research areas short answer assessment, authorship attribution, and plagiarism detection.

The experiments confirmed the hypothesis formulated in the introduction that these areas share a similar methodology in terms of frameworks, tasks, and features. It was shown that semantics-based features modeling aspects of content, especially robust character-based features, were most effective for both short answer assessment and plagiarism detection, and that the most informative features for both corpora were surprisingly similar. The experiments also made evident that already rather simple features can yield reasonable results for these tasks.

Both research hypotheses formulated in section 5 could be confirmed, respectively the null hypothesis could be rejected: features from authorship attribution yielded significant improvements for the task of learner language assessment, and features from learner language assessment yielded significant improvements for the task of plagiarism detection. However, it has to be noted that not all features are strictly task-specific, and also applicable to other NLP tasks.

A comparison with related work showed that the results are comparable to current state-of-the-art approaches, although there is still room for improvement. Future work therefore will explore the usage of more features, more elaborate machine learning algorithms, and automatic feature selection techniques. In addition, more corpora from either domain will be used to obtain a broader evaluation perspective. Especially stylistic features modeling for example stopword patterns as well as longest common subsequence features are hypothesized to be beneficial for the task of plagiarism detection since they model stylistic rather than semantic properties.

Acknowledgments

I would like to thank the three anonymous reviewers and Ramon Ziai for their insightful comments.

References

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, Jill Burstein, and Rachele De Felice, editors, *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 107–115, Columbus, Ohio.

Jason Baldridge. 2005. The OpenNLP Project. *URL: http://opennlp. apache. org/index. html,(accessed 25 August 2015)*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Paul Clough and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, 2007. *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, Tilburg, The Netherlands, July 11. Version 6.0.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

David Gale and Lloyd S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly*, 69:9–15.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3):251–270.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, pages 94–103, Montreal.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Andrea Horbach, Alexis Palmer, and Manfred Pinkal. 2013. Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 286–295, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011a. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation*, 21(4):355–369.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011b. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, July. ACL.

Joakim Nivre, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(1):1–41.

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Ulrike Pado and Cornelia Kiefer. 2015. Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA*, page 43.

Björn Rudzewitz. 2015. Alignment Weighting for Short Answer Assessment. Bachelor's thesis, University of Tübingen.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502, Freiburg, Germany.

Daniel Bär Torsten Zesch and Iryna Gurevych. 2012. Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING*, volume 1, pages 167–184.

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, pages 159–168, Dublin, Ireland. COLING, Association for Computational Linguistics.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, pages 190–200, Montreal.

Sven Meyer Zu Eissen and Benno Stein. 2006. Intrinsic plagiarism detection. In *Advances in Information Retrieval*, pages 565–569. Springer.