Semantics-based automated essay evaluation

A dissertation presented

by

Kaja Zupanc

to

The Faculty of Computer and Information Science

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in the subject of

Computer and Information Science

Ljubljana, 2018

APPROVAL

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

— Kaja Zupanc —

May 2018

The submission has been approved by

dr. Janez Demšar

Professor of Computer and Information Science

EXAMINER

dr. Matjaž Kukar

Associate Professor of Computer and Information Science

EXAMINER

dr. Richard Johansson

Associate professor of Computer and Information Science

EXTERNAL EXAMINER

University of Gothenburg, Sweden

dr. Zoran Bosnić

Associate Professor of Computer and Information Science

ADVISOR

PREVIOUS PUBLICATION

I hereby declare that the research reported herein was previously published/submitted for publication in peer reviewed journals or publicly presented at the following occasions:

[1] K. Zupanc, and Z. Bosnić. Automated essay evaluation augmented with semantic coherence measures. In R. Kumar, editor, *Proc. of ICDM 2014*, pages 1133-1138, Los Alamitos (CA), 2014. IEEE. doi: 10.1109/ICDM.2014.21

[2] K. Zupanc, and Z. Bosnić. Advances in the field of automated essay evaluation. *Informatica*, 39(4):383–396, 2015.

[3] K. Zupanc, and Z. Bosnić. Automated essay evaluation with semantic analysis. *Knowledge-based systems*, 120:118-132, 2017. doi: 10.1016/j.knosys.2017.01.006

[4] K. Zupanc, M. Savić, Z. Bosnić, and M. Ivanović. Evaluating coherence of essays using sentence-similarity networks. In B. Rachev, and A. Smrikarov, editors, *Proc. of CompSysTech 2017*, volume 1369 of *ACM International Conference Proceeding Series*, pages 65-72, New York (NY), 2017. ACM. doi: 10.1145/3134302.3134322

[5] K. Zupanc, and Z. Bosnić. Improvement of automated essay grading by grouping similar graders. Under review, 2018.

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Ljubljana.

This dissertation is a result of doctoral research, in part financed by the European Union, European Social Fund and the Republic of Slovenia, Ministry for Education, Science and Sport in the framework of the Operational programme for human resources development.

University *of Ljubljana*
Faculty *of Computer and Information Science*
Kaja Zupanc
Semantics-based automated essay evaluation

ABSTRACT

Automated essay evaluation (AEE) is a widely used practical solution for replacing time-consuming manual grading of student essays. Automated systems are used in combination with human graders in different high-stake assessments, as well as in classrooms. During the last 50 years, since the beginning of the development of the field, many challenges have arisen in the field, including seeking ways to evaluate the semantic content, providing automated feedback, determining reliability of grades, making the field more "exposed", and others. In this dissertation we address several of these challenges and propose novel solutions for semantic based essay evaluation.

Most of the AEE research has been conducted by commercial organizations that protect their investments by releasing proprietary systems where details are not publicly available. We provide comparison (as detailed as possible) of 20 state-of-the-art approaches for automated essay evaluation and we propose a new automated essay evaluation system named SAGE (Semantic Automated Grader for Essays) with all the technological details revealed to the scientific community.

Lack of consideration of text semantics is one of the main weaknesses of the existing state-of-the-art systems. We address the evaluation of essay semantics from perspectives of essay coherence and semantic error detection. Coherence describes the flow of information in an essay and allows us to evaluate the connections between the discourse. We propose two groups of coherence attributes: coherence attributes obtained in a highly dimensional semantic space and coherence attributes obtained from a sentence-similarity networks. Furthermore, we propose the Automated Error Detection (AED) system and evaluate the essay semantics from the perspective of essay consistency. The system detects semantic errors using information extraction and logic reasoning and is able to provide semantic feedback for the writer. The proposed system SAGE achieves significantly higher grading accuracy compared with other state-of-the-

art automated essay evaluation systems.

In the last part of the dissertation we address the question of reliability of grades. Despite the unified grading rules, human graders introduce bias into scores. Consequently, a grading model has to implement a grading logic that may be a mixture of grading logics from various graders. We propose an approach for separating a set of essays into subsets that represent different graders, which uses an explanation methodology and clustering. The results show that learning from the ensemble of separated models significantly improves the average prediction accuracy on artificial and real-world datasets.

Keywords: automated scoring, essay evaluation, natural language processing, semantic attributes, coherence, semantic feedback

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*
Kaja Zupanc
Semantično usmerjeno avtomatsko ocenjevanje esejev

POVZETEK

Avtomatsko ocenjevanje esejev predstavlja praktično rešitev za številne težave, povezane s časovno zahtevnim ročnim ocenjevanjem. Avtomatizirani sistemi se uporabljajo v kombinaciji s človeškimi ocenjevalci pri številnih standardiziranih testih, vse več pa tudi v učilnicah. V zadnjih 50 letih, od začetka razvoja področja, so se pojavili številni izzivi, vključno z iskanjem pristopov za ocenjevanje semantične vsebine, zagotavljanjem avtomatskih povratnih informacij, določanjem zanesljivosti ocen, težnjo po dostopnosti podrobnosti delovanja sistemov in s tem odprtosti področja, in drugi. V pričujoči disertaciji obravnavamo te izzive in predlagamo nove rešitve za semantično usmerjeno avtomatsko ocenjevanje esejev.

Eden od glavnih problemov sistemov za avtomatsko ocenjevanje esejev je problem ocenjevanja semantične pravilnosti besedila. V disertaciji obravnavamo ocenjevanje semantike besedila z različnimi pristopi: ocenjevanje koherence esejev in zaznavanje semantičnih napak. Koherenca opisuje pretok informacij v eseju in nam omogoča, da ocenimo povezanost besedila. Predlagamo dve skupini atributov za ocenjevanje koherence: atributi, pridobljeni v visoko dimenzionalnem semantičnem prostoru, in atributi, pridobljeni iz omrežij stavčne podobnosti. Poleg tega predlagamo sistem za avtomatsko odkrivanje napak, ki nam pomaga oceniti semantiko eseja z vidika doslednosti. Sistem zaznava semantične napake z uporabo ekstrakcije informacij in logičnega sklepanja ter zagotavlja povratno semantično informacijo. Predlagani sistem SAGE (*Semantic Automated Grader for Essays*) dosega višjo napovedno točnost v primerjavi z drugimi sodobnimi sistemi za avtomatsko ocenjevanje esejev.

V zadnjem delu disertacije se posvečamo vprašanju zanesljivosti ocen. Kljub poenotenim kriterijem za človeške ocenjevalce, ocenjevalci vnašajo pristranskost v rezultate. Zato mora napovedni model uporabiti napovedno logiko, ki je lahko mešanica ocenjevalne logike različnih ocenjevalcev. Predlagamo pristop za ločevanje množice esejev v

podmnožice, ki predstavljajo različne ocejevalce, kjer uporabimo metodologijo razlage napovedi in gručenje. Rezultati kažejo, da učenje na ansamblu ločenih modelov bistveno izboljša povprečno točnost napovedi na umetnih in realnih podatkovnih množicah.

Ključne besede: avtomatsko ocenjevanje, evalvacija esejev, procesiranje naravnega jezika, semantični atributi, skladnost, semantična povratna informacija

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Zoran Bosnić. You have a remarkable gift of patience! I am thankful for all the remarks you had on my work and I am even thankful for your never-ending tendency to improve my writing skills. You believed in me from the beginning of my PhD and I will always be grateful for that!

I would like to thank my thesis committee: prof. Janez Demšar, assoc. prof. Matjaž Kukar, and assoc. prof. Richard Johansson, for their insightful comments and suggestions that certainly improved my thesis. I would especially like to thank assoc. prof. Johansson for agreeing to serve as an external examiner of my dissertation.

My special thanks goes to prof. Jesse Davis for being my advisor during my research stay at KU Leuven, and for all the discussions we had even after my time in Leuven. I would also like to thank prof. Mirjana Ivanović for hosting me at the University of Novi Sad.

My lab has not only been a source of good advice, but also a source of friendship. You helped me survive the PhD journey! Erik, thank you for all the on and off-topic conversations, funny moments, and bad jokes that put a smile on my face!

This journey would not have been possible without the support of my family. Thank you for supporting me throughout writing this thesis and through my life in general. I am especially grateful to my mum for all the emotional support: Mami, hvala!

Thanks to all my friends for sharing my happiness when starting my PhD and following with encouragement and entertainment when it seemed too difficult to be completed.

Last but not least, I would like to thank Ben, for staying loving, positive, and supporting even at the hardest times. Dankjewel!

— Kaja Zupanc, Ljubljana, May 2018.

CONTENTS

I

Introduction

1.1 Motivation

Essays are short literary compositions on a particular subject (also referred to as prompt-specific essays), usually in prose, and generally analytic, speculative, or interpretative in nature. Essays are considered to be the most useful tool to assess learning outcomes, giving students an opportunity to demonstrate their range of skills and knowledge, including higher-order thinking skills, such as synthesis and analysis [Valenti et al., 2003]. However, grading students' essays is a time-consuming, labor-intensive, and expensive activity for educational institutions. Since teachers are burdened with hours of grading of written assignments, they assign less essay writing tasks, thereby limiting the needed experience to reach the writing excellency. This contradicts the aim to make students better writers, for which they need to rehearse their skill by writing as much as possible [Page, 1966].

A practical solution to many problems associated with manual grading is to have an automated system for essay evaluation. Automated essay evaluation (AEE) is the process of evaluating and scoring written essays via computer programs [Shermis and Burstein, 2003]. For teachers and educational institutions, AEE represents not only a tool to assess learning outcomes, but also helps save time, effort, and money without lowering the quality of evaluation.

AEE is a multi-disciplinary field that incorporates research from computer science, cognitive psychology, educational measurement, linguistics, and writing research [Shermis et al., 2013]. Computer scientists are developing attributes and are implementing AEE systems, writing scientists and teachers are providing constructive criticisms to the development, and cognitive psychologists expert opinion is considered when modelling the attributes. Psychometric evaluations provide crucial information about the reliability and validity of the systems, as well.

The field has been developing since the 1960s when Ellis Batten Page and his colleagues proposed the first automated essay scoring (AES) system [Page, 1966]. The system was using basic measures to approximate features of interest and thus describe the quality of an essay. By the 1990s, the progress in the natural language processing (NLP) field encouraged researchers to apply new computational techniques to automatically extract essay writing quality measures. In the last decade, this automated process became the preferred way of grading in many low-stake assessment in classrooms as well as in high-stake assessment as standardized tests. AEE systems can also

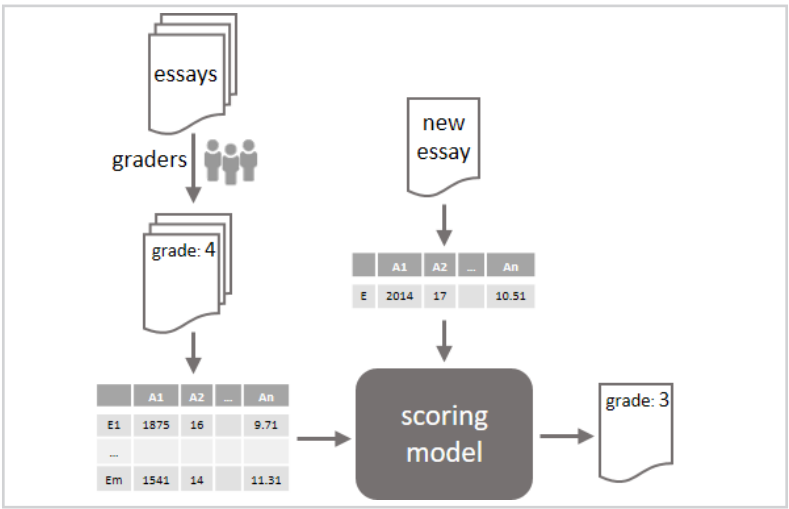

Figure 1.1

Illustration of an automated essay evaluation: A set of essays is pre-scored by human graders and a computer program extracts attributes representing each essay to build a training set. The set is used to develop the scoring model. This scoring model is used to assign the scores to new, ungraded essays.

be used in all other application areas of text mining, where the content of the text needs to be graded or prioritized, such as: written applications, cover letters, scientific papers, e-mail classification etc.

Framework of operation (shown in Figure 1.1) is a common property of majority of the AEE systems: Systems use a substantially large set of prompt-specific essays (i.e. set of essays on the same topic) assessed by expert human graders. A computer program extracts a set of attributes for each of these essays to construct the learning set. This set is used to build the scoring model of the AEE system. Using this model, the AEE system assigns scores to new, ungraded essays. The performance of the scoring model is typically validated by calculating how well the scoring model replicated the scores assigned by the human expert graders using metrics such as quadratic weighted Kappa and exact agreement [Fazal et al., 2011].

Throughout the development of the field, several different names have been used for the field interchangeably. The terms automated essay scoring (AES) and automated essay grading (AEG) slowly became replaced with the term automated writing evaluation (AWE) or automated essay evaluation (AEE). The term *evaluation* within the name (AWE, AEE) surfaced to use because the automated process enables students to receive constructive feedback about their writing.

1.2 Problem description

The field of automated essay evaluation is currently (as of 2018) focusing on three main open challenges:

1. *Weak consideration of text semantics.*
   The main weakness of existing AEE systems is that they consider text semantics very weakly and focus mostly on its syntax. Although the details of the majority of the systems have never been announced publicly, we can still deduce from the literature that they mostly perform syntax and shallow content measurements (calculating similarity between texts) and neglect the semantics. To analyse semantics, the state-of-the-art systems use latent semantic analysis (LSA) [Landauer et al., 1998], latent Dirichlet allocation (LDA) [Kakkonen et al., 2008], and content vector analysis (CVA) [Attali, 2011]. To measure the coherence of essays' content, LSA [Foltz et al., 1998; Foltz, 2007], random indexing [Higgins et al., 2004], an entity-based approach [Burstein et al., 2010], and complex network representation [Antiqueira et al., 2007; Ke et al., 2016] have been used. However, only two existing systems [Gutierrez et al., 2014; Brent et al., 2010] use approaches that partially check for consistency of the statements in the essays. Despite the efforts, the latter systems are not automatic, as they require manual interventions from the user.

2. *Reproduction of biased human grades.*
   In the last years, many researchers [Bejar, 2011; Attali, 2013; Williamson et al., 2012] have debated that accurately reproducing the human graders is no longer the main goal of AEE systems. Researchers in the field of Automated Essay Evaluation (AEE) often consider expert human graders as unmistakeable and objective, but scoring essays relies heavily on human judgement. In reality, human graders are inconsistent and unreliable. Biased scoring is thought to be due to various aspects of reader characteristics (e.g., rating experiences), reader psychology (factors that occur internally to the reader), and rating environment (including pressure) [Bridgeman, 2013]. Scores are therefore subjective and influenced by *grader effects*. That is, scores can be affected by factors, such as bias (strictness, leniency) and (un)reliability (non-systematic error) of the grader. Systematic and non-systematic human errors introduce subjective variance into

scores and therefore impact their validity [Lottridge et al., 2013].
It is desirable that the AEE systems can recognize certain types of errors independently of human raters, including syntactic errors, and offer automated feedback on correcting these errors. In addition, the systems shall also provide global feedback on content and development. The current limitation of the feedback is that its content is limited to the syntactic aspect of the essay while neglecting the semantic aspects. Exceptions are systems [Gutierrez et al., 2014; Brent et al., 2010] that include semantic evaluation of the content, but are not automatic.

3. *Lack of standards and good practice (due to predominance of proprietary systems).* In the past, one of the main obstacles to achieve progress in this area was the lack of open-source AEE systems, which would provide insight into their grading methodology. Namely, most of the AEE research and development has been conducted by commercial or non-profit organisations that have protected their investments by restricting access to the technological details. The first scoring engine to be made available publicly was Rudner's Bayesian Essay Test Scoring sYstem (BETSY) [Rudner and Liang, 2002]. This was a preliminary investigation and authors never continued with their work. More recently, Mayfield and Rosé released LightSIDE [Mayfield and Penstein-Rosé, 2010], an easy-to-use automated evaluation engine with both compiled and source code publicly available. LightSIDE made a very important contribution to the field of AEE by publicly providing the source code. However, the used methodology to predict the final grade is very basic. Aside from these two systems, there were several other attempts in the last couple of years to make the field more transparent, including publishing the Handbook of Automated Essay Evaluation [Shermis and Burstein, 2013].

1.3 Scientific contributions

In the light of the preceding discussion about the existing models' weaknesses, the following scientific contributions are presented in this dissertation:

1. *New semantic attributes for evaluating semantic coherence of the text.* We propose two groups of coherence attributes: spatial attributes and network attributes. We develop the first group of attributes by observing semantic changes within

the flow of the text. To achieve this contribution, we complete the following sub-tasks: optimally choosing sequential parts of the essay to be observed, building the semantic space by transforming essay parts using a variation of TF-IDF vectors, and defining several metrics for describing variability of the essay parts in that space. We obtain the second group of the attributes using sentence-similarity networks. Within the network we represent each sentence as a node and use weights to represent similarity between sentences. We derive different structural metrics from such networks to propose novel coherence attributes. Furthermore, we develop a new automated essay grading system that uses the novel coherence attributes and perform systematic analysis of its performance compared with all available state-of-the-art systems. On average the system significantly improves the classification accuracy due to the novel attributes.

2. *Methodology for cross-referencing facts in text with external fact sources.* This contribution yields additional semantic attributes that monitor the content of the essay and detect the sentences that contradict the truth. We develop *automated error detection system* (AED), a novel methodology that discovers semantic errors and provides a comprehensive feedback. We define an approach to automatically transform essay text into independent representation (ontology) and compare it to a representation of common sense knowledge from the external sources in a form of ontology. We enhance the developed automated essay evaluation system with new attributes and the output of the AED system, i.e., automated semantic feedback. We also published all the technological details of the system.

3. *Methodology for detection of different graders.* We propose a novel methodology for separation of the original dataset that contains scores given by several different graders into smaller subsets. We aim these smaller subsets to contain only essays that were graded by the same grader. To differentiate between different graders we use the explanation methodology and clustering, which enables us to detect different dependencies (grading logics) between essays' attributes and its score. In our experiments we show that the single model learned on resolved scores performs worse on the average than the ensemble of models that represents individual different graders. Our results show that the approach is able to detect different graders using unsupervised learning and obtain better predictive performance on the ensemble of models. We augment the proposed

AEE system with the methodology and we significantly improve the prediction accuracy.

1.4 Dissertation overview

The main aim of this dissertation is to build a semantic automated essay evaluator that would not only improve the grading accuracy but would be capable of providing understandable feedback about essay's semantics. Thus, we first build a basic system and then upgrade it through the subsequent chapters and sections.

In Chapter 2 we first describe several subfields of the related work that are relevant to our research. Most importantly, we compare the existing state-of-the-art systems that represent a basis for our research. Through the Chapter 3 we build a new automated essay evaluation system and upgrade it with novel coherence and consistency attributes. We conclude the chapter with the proposal of the automated error detection system as part of the SAGE (Semantic Automated Grader for Essays) that provides semantic feedback for students. In Chapter 4 we describe the methodological aspects of the proposed system SAGE. We perform several experiments to demonstrate the performance of the system. Among others, we compare the system with nineteen other state-of-the art AEE approaches. Chapter 5 proposes a new methodology for grouping similar graders into subsets which reflects in an improved grading prediction accuracy and can be used on any AEE system. Chapter 6 draws conclusions.

2

Automated essay evaluation

We approached the open challenges of the Automated Essay Evaluation (AEE) field from different problem areas that we highlight in the following subsections. We start with the overview and comparison of the existing AEE systems and continue with the description of the state-of-the-art approaches for measuring coherence of an essay. Next, we review the relevant fields for detection of semantic errors in an essay and we conclude with the description of the research on biased scoring.

2.1 Automated essay evaluation systems

In 1966, the high school English teacher E. Page proposed the first automated system for grading student essays [Page, 1966]. He saw the system as a solution to reducing hours of manually grading student essays. In 1973 [Ajay et al.] he and his colleagues had enough hardware and software at their disposal to implement the first AEE system under the name Project Essay Grade. The first results were characterized as remarkable as the system's performance had more steady correlation with human graders than the performance of two trained human graders. Despite its impressive success at predicting teachers' essay ratings, the early version of the system received only limited acceptance in writing and education community. The availability of necessary tools (home computers, Internet, computational techniques for automatically extracting measures of writing quality, ...) was poor and the society criticised the idea of displacing human graders [Shermis et al., 2013].

By the 1990s, with the widespread of the Internet, natural language processing tools, e-learning systems, and statistical methods, the AEE became a support technology in education. Nowadays, the AEE systems are used in combination with human graders in different high-stake assessments such as the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), Graduate Management Admissions Test (GMAT), Scholastic Aptitude Test (SAT), American College Testing (ACT), Test of English for International Communication (TOEIC), Analytic Writing Assessment (AWA), No Child Left Behind (NCLB) and Pearson Test of English (PTE). Furthermore, some of them also act as a sole grader.

In the past, one of the main obstacles to achieve progress in this area was lack of open-source AEE systems, which would allow insight into their grading methodology. In the following we present the known characteristics of the majority of proprietary AEE systems developed by commercial organizations as well as two publicly-available systems and approaches proposed by the academic community. We conclude this sec-

tion with a comparison of described systems and their known characteristics.

- *Project Essay Grade (PEG)*
  PEG is a proprietary AES system developed at Measurement Inc. [Page, 1966]. It was first proposed in 1966, and in 1998 a web interface was added [Shermis et al., 2001]. The system scores essays through measuring *trins* and *proxes*. A *trin* is defined as an intrinsic higher-level variable, such as punctuation, fluency, diction, grammar, etc., which as such cannot be measured directly and has to be approximated by means of other measures, called *proxes*. For example, the trin *punctuation* is measured through the proxes *number of punctuation errors* and *number of different punctuations used*. The system uses regression analysis to score new essays based on a training set of 100 to 400 essays [Page, 1994].

- *e-rater*
  E-rater is a proprietary automated essay evaluation and scoring system developed at the Educational Testing Service (ETS) in 1998 [Burstein et al., 1998]. E-rater identifies and extracts several attribute classes [Attali and Burstein, 2006; Burstein et al., 2004]: (1) grammatical errors (e.g. subject-verb agreement errors), (2) word usage errors (e.g. their versus there), (3) errors in writing mechanics (e.g. spelling), (4) presence of essay-based discourse elements (e.g. thesis statement, main points, supporting details, and conclusions), (5) development of essay-based discourse elements, (6) style weaknesses (e.g. overly repetitious words), (7) two content vector analysis (CVA)-based attributes to evaluate topical word usage, (8) an alternative, differential word use content measure, based on the relative frequency of a word in high scoring versus low-scoring essays, (9) two attributes to assess the relative sophistication and register of essay words, and (10) an attribute that considers correct usage of prepositions and collocations (e.g. powerful computer vs. strong computer), and variety in terms of sentence structure formation. The system uses regression modelling to assign a final score to the essay [Burstein et al., 2013a]. E-rater also includes detection of essay similarity and advisories that point out if an essay is off topic, has problems with discourse structure, or includes large number of grammatical errors [Higgins et al., 2006].

- *Intelligent Essay Assessor (IEA)*
  In 1998 the Pearson Knowledge Technologies (PKT) developed Intelligent Essay Assessor (IEA). The system is based on the Latent Semantic Analysis (LSA), a machine-learning method that acquires and represents knowledge about meaning of words and documents by analysing large bodies of natural text [Landauer et al., 2000]. IEA uses LSA to derive attributes describing content, organization, and development-based attributes of writing. Along with LSA, IEA also uses NLP-based measures to extract attributes measuring lexical sophistication, grammatical, mechanical, stylistic, and organizational aspects of essays. The system uses approximately 60 attributes to measure the above aspects within essays: content (e.g. LSA essay semantic similarity, vector length), lexical sophistication (e.g. word maturity, word variety, and confusable words), grammar (e.g. n-gram attributes, grammatical errors, and grammar error types), mechanics (e.g. spelling, capitalization, and punctuation), style, organization, and development (e.g. sentence-sentence coherence, overall essay coherence, and topic development). IEA requires a training with a representative sample (between 200 and 500) of human-scored essays.

- *IntelliMetric*
  IntelliMetric was designed and first released in 1999 by Vantage Learning as a proprietary system for scoring essay-type, constructed response questions [Schultz, 2013]. The system analyses more than 400 semantic-, syntactic-, and discourse-level attributes to form a composite sense of *meaning*. These attributes can be divided into two major categories: content (discourse/rhetorical and content/concept attributes) and structure (syntactic/structural and mechanics attributes). The content attributes evaluate the covered topic, the breadth of content, support for advanced concepts, logic of discourse, as well as cohesiveness and consistency in purpose and main idea; whereas structure attributes evaluate grammar, spelling, capitalization, sentence completeness, punctuation, syntactic variety, sentence complexity, usage, readability, and subject-verb agreement [Schultz, 2013]. The system uses multiple predictions (called judgements) based on multiple mathematical models, including linear analysis, Bayesian approach, and LSA to predict the final score and combines the models into a single final essay score [Rudner et al., 2006]. Training Intellimetric requires a sample of

at least 300 human-scored essays. IntelliMetric uses Legitimatch technology to identify responses that appear off topic, are too short, do not conform to the expectations for edited American English, or are otherwise inappropriate [Schultz, 2013].

- *Bookette*
  Bookette [Rich et al., 2013] was designed by California Testing Bureau (CTB) and became operational in classroom settings in 2005 and in large-scale testing settings in 2009. Bookette uses NLP to derive about 90 attributes describing student-produced text. Combinations of these attributes describe traits of effective writing: organization, development, sentence structure, word choice/grammar usage, and mechanics. The system uses neural networks to model expert human grader scores. Bookette can build prompt-specific models as well as generic models that can be very useful in classrooms for formative purposes. Training Bookette requires a set (from 250 to 500) of human-scored essays. The system provides feedback on students writing performance that includes both holistic feedback and feedback at the trait level including comments on the grammar, spelling, and writing conventions at the sentence level [Rich et al., 2013].

- *CRASE*
  Pacific Metrics proprietary automated scoring engine, CRASE [Lottridge et al., 2013], moves through three phases of the scoring process: identifying inappropriate attempts, attribute extraction, and scoring. The attribute extraction step is organized around six traits of writing: ideas, sentence fluency, organization, voice, word choice, conventions, and written presentation. The system analyses a sample of already-scored student responses to produce a model of the graders' scoring behaviour. CRASE is a Java-based application that runs as a web service. The system is customizable with respect to the configurations used to build machine learning models as well as the blending of human and machine scoring (i.e. deriving hybrid models) [Lottridge et al., 2013]. Application also produces text-based and numeric-based feedback that can be used to improve the essays.

- *AutoScore*
  AutoScore is a proprietary AEE system designed by the American Institute for

Research (AIR). The system analyses measures based on concepts that discriminate between high- and low- scored papers, measures that indicate the coherence of concepts within and across paragraphs, and a range of word-use and syntactic measures. Details about the system were never published, however, the system was evaluated in [Shermis and Hamner, 2013].

- *Lexile Writing Analyzer*
  The Lexile Writing Analyzer is a part of The Lexile Framework for Writing [Smith, 2009] developed by MetaMetrics. The system is score-, genre-, prompt-, and punctuation-independent and utilizes the Lexile writer measure, which is an estimate of student's ability to express language in writing, based on factors related to semantic complexity (the level of words used) and syntactic sophistication (how the words are written into sentences). The system uses a small number of attributes that represent approximations for writing ability. Lexile perceives writing ability as an underlying individual trait. Training phase is not needed since a vertical scale is employed to measure student essays [Smith et al., 2014].

- *SAGrader*
  SAGrader is an online proprietary AEE system developed by IdeaWorks, Inc. [Brent and Townsend, 2006]. The system was first known under the name Qualrus. SAGrader blends a number of linguistic, statistical, and artificial intelligence approaches to automatically score the essay. The operation of the SAGrader is as follows: The instructor first specifies a task in a prompt. Then the instructor creates a rubric identifying the "desired features" – key elements of knowledge (set of facts) that should be included in a good response, along with relationships among those elements – using a semantic network (SN). Fuzzy logic (FL) permits the program to detect the features in the students' essays and compare them to desired ones. Finally, an expert system scores student essays based on the similarities between the desired and observed features [Brent et al., 2010]. Students receive immediate feedback indicating their scores along with the detailed comments indicating what they did well and what needs further work.

- *OBIE based AEE System*
  The AEE system proposed by Gutierrez et al. [2012, 2013, 2014] provides both, scores and meaningful feedback, using ontology-based information extraction (OBIE). The system uses logic reasoning to detect errors in a statement from an essay. The system first transforms text into a set of logic clauses using open information extraction (OIE) methodology and incorporates them into domain ontology using manually selected vocabulary mapping. The system determines if these statements contradict the ontology and consequently the domain knowledge. This method considers incorrectness as inconsistency with respect to the domain. Logic reasoning is based on the description logic (DL) and ontology debugging [Gutierrez et al., 2014].

- *Bayesian Essay Test Scoring sYstem (BETSY)*
  The first scoring engine to be made available publicly was Rudner's Bayesian Essay Test Scoring sYstem (BETSY) [Rudner and Liang, 2002]. BETSY uses multinomial or Bernoulli Naïve Bayes models to classify texts into different classes (e.g. pass/fail, scores A-F) based on content (e.g. word uni-grams and bi-grams) and style attributes (e.g. sentence length). Classification is based on assumption that each attribute is independent of another. BETSY worked well only as a demonstration tool for a Bayesian approach to scoring essays. It remained a preliminary investigation as the authors never continued with their work.

- *LightSIDE*
  In 2010, Mayfield and Rosé released LightSIDE [Mayfield and Penstein-Rosé, 2010], an automated evaluation engine with both compiled and source code publicly available. LightSIDE is designed as a tool for non-experts to effectively use text mining technology for a variety of purposes, including essay assessment. It allows choosing the set of attributes and algorithm to build prediction model (e.g. linear regression, Naïve Bayes, linear support vector machines) [Mayfield and Rosé, 2013]. The set of attributes is mainly focused on n-grams, POS tags, and "counting" attributes. However, the system allows users to manually input the code for new attributes.

- *Use of Syntactic and Shallow Semantic Tree Kernels for AEE*
  Chali and Hasan [2013] exposed the major limitation of LSA - it only retains the frequency of words by disregarding the word sequence and the syntactic and semantic structure of texts. They proposed the use of syntactic and shallow semantic tree kernels for grading essays as a substitute to LSA. The system calculates the syntactic similarity between two sentences by parsing the corresponding sentences into syntactic trees and measuring the similarity between the trees. Shallow Semantic Tree Kernel (SSTK) method allows to match portions of a semantic trees. The SSTK function yields the similarity score between a pair of sentences based on their semantic structures.

- *A Ranked-based Approach to AEE*
  Chen et al. [2012] consider the problem of AEE as a ranking problem instead of classification or regression problem. Ranking algorithms are a family of supervised learning algorithms that automatically construct a ranking model of the retrieved essays. They consider the following three groups of attributes: term usage, sentence quality, and content fluency and richness. Authors showed that their approach outperforms other classical machine learning techniques.

- *Neural Essay Assessor*
  Taghipour and Ng [2016] developed an approach based on recurrent neural networks. The system learns the relations between an essay and its assigned score automatically, without any feature engineering. It encodes the information required for essay grading and learns the complex patterns in the data through non-linear neural layers. The neural network is based on long-term memory networks.

- *AES using neural networks*
  Alikaniotis et al. [2016] introduced a model that forms word representations (embeddings) by learning the local linguistic environment of each word as well as the extent to which a specific word contributes to an essay's score. The system uses the long short-term memory recurrent neural networks [Hochreiter and Schmidhuber, 1997] to represent the meaning of essays.

- *AEG using Memory Networks*
  Zhao et al. [2017] proposed a generic model using memory networks inspired

by a network's capability to store rich representations of data and reason over that data in memory. Their model is based on the idea that with enough graded samples for each score in the rubric, such samples can be used to grade future work that is found to be similar. For each possible score in the rubric, a student response graded with the same score is collected. These selected responses represent the grading criteria specified in the rubric and are stored in the memory component. The model is trained on these data with the rest of the student responses in a supervised learning manner to compute the relevance between the representation of an ungraded response and that of each sample.

- *OzEgrader*
  OzEgrader is an Australian AES system proposed by Fazal et al. [2011]. Grading process considers different aspects of content and style: audience, text structure, character and setting, paragraphing, vocabulary, sentence structure, punctuation, spelling, cohesion and ideas. Techniques such as POS tagging, named entity recognition, artificial neural networks, and fuzzy regression are employed in order to model linear or non-linear relationships between attributes and the final score. The system also includes the methodology for noise reduction in the essay dataset.

- *AEE using Generalized LSA*
  Islam and Hoque [2012] developed an AEE system using Generalized Latent Semantic Analysis (GLSA) which makes *n-gram by document* matrix instead of *word by document* matrix as used in LSA. The system uses the following steps in grading procedure: preprocessing of the training essays, stopword removal, word stemming, selecting the n-gram index terms, *n-gram by document* matrix creation, computation of the singular value decomposition (SVD) of *n-gram by document* matrix, dimensionality reduction of the SVD matrices, and computation of the similarity score. The main advantage of GLSA is observance of word order in sentences.

- *AEE using Multi-classifier Fusion*
  Bin and Jian-Min [2011] proposed an approach to AEE using multi-classifier Fusion. The system first represents each essay by the vector space model and removes stopwords. Then it extracts the attributes describing content and lin-

guistic from the essays in the form of attribute vector. Three approaches including document frequency (DF), information gain (IG) and chi-square statistic (CHI) are used to select attributes by some predetermined thresholds. The system classifies an essay to an appropriate category using different classifiers, such as naïve Bayes, k-nearest neighbours and support vector machine. Finally, the ensemble classifier is combined by those component classifiers.

- *Markit*
  Markit [Williams and Dreher, 2004] is a proprietary AEE system developed by Blue Wren Software Pty Ltd. The system is capable of running on typical desktop PC platforms. It requires comprehensive knowledge in a form of one model (exemplary) answer against which the student essays are compared. A student essay is processed using a combination of NLP techniques to build the corresponding propriety knowledge representation. Pattern matching techniques (PMT) are then employed to ascertain the proportion of the model answer knowledge that is present in the student's answer, and a score assigned accordingly.

- *PS-ME*
  The Paperless School proprietary AEE system was designed primarily for day-to-day, low-stake testing of essays. The student essay is compared against each relevant master text to derive a number of parameters which reflect knowledge and understanding as exhibited by the student. When multiple master texts are involved in the comparison, each result from an individual comparison gets a weight that could be negative in the case of a master text containing common mistakes. The individual parameters computed during the analysis phase are then combined in a numerical expression to yield the assignments' score and used to select relevant feedback comments from a comment bank [Mason and Grove-Stephenson, 2002].

- *Schema Extract Analyse and Report (SEAR)*
  Christie [1999] proposed a software system Schema Extract Analyse and Report (SEAR), which provides the assessment both of style and content. The methodology adopted to assess style is based on a set of common metrics as used by other AES systems. For content assessment the system uses two measures: us-

age and coverage. Using content schema system measures how much of each essay is included in schema (usage) and how much of schema is used by the essay (coverage).

Comparison of the state-of-the-art systems

In order to compare technical, methodological, and structural characteristics of the state-of-the-art systems, we present their comparison in Table 2.1 using the following main criteria:

- *Type of attributes*: The attributes describing the quality of an essay can roughly be divided into three groups: *style*, *content* and *semantic* attributes. Style attributes focus on lexical sophistication, grammar and mechanics (spelling, capitalization, and punctuation). Content attributes shallowly describe semantics of an essay and are based on comparing an essay with source text and other already graded essays. Semantic attributes are based on verifying the correctness of content meaning.

- *Methodology*: Different systems use various approaches to extract attributes from essays. The most widely used methodology is based on NLP. Systems focusing on content mostly use Latent Semantic Analysis (LSA) - a machine learning method that analyses related concepts between a set of documents and the contained terms. LSA assumes that words with similar meaning occur in similar parts of text. To evaluate content, systems also use pattern matching techniques (PMT) and extensions to LSA such as Generalized Latent Semantic Analysis (GLSA) (which uses an *n-gram-by-document* matrix instead of a *word-by-document* matrix) and improvement that considers semantics by means of the syntactic and shallow semantic tree kernels. For verifying the correctness and consistency of content, approaches such as (Open) Information Extraction ((O)IE), Semantic Networks (SN), Ontologies, Fuzzy Logic (FL), and Description Logic (DL) are used.

- *Prediction model*: The majority of the systems use machine learning algorithms (usually regression modelling) to predict the final grade. An alternative is to use: the Lexile measure - an estimate of student's ability to express language in writing based on semantic complexity (level of expressing) and syntactic sophistication

(how the words are combined into sentences); cosine similarity; and rule-based expert systems.

Table 2.1 provides a comparison of characteristics for the majority of AEE systems and approaches, including proprietary (non-public) systems, two publicly available systems, approaches proposed by the academic community, and our proposed system named *SAGE*.

2.2 Measuring coherence of the text

Coherence is a concept that describes the flow of information from one part of discourse to another and ranges from lower level cohesive elements such as coreference, causal relationship, and connectives, up to higher level elements that evaluate connections between the discourse and reader's mental representation of it [Foltz, 2007].

Existing systems measure coherence in noisy text with different supervised and unsupervised approaches. The unsupervised approaches usually measure lexical cohesion, i.e. repetition of words and phrases in an essay. Foltz, Kintsch, and Landauer [Foltz et al., 1998; Foltz, 2007] assume that coherent texts contain a high number of semantically related words and measure coherence as a function of semantic relatedness between adjacent sentences. Relatedness can be computed using LSA without employing syntactic or other annotations. Hearst [1997] subdivides texts into multi-paragraph units that represent subtopics and identifies patterns of lexical co-occurrence and distribution, i.e. identifying repetition of vocabulary across adjacent sentences.

The supervised learning approaches require annotated data (graded essays). They focus on occurrences of discourse elements (e.g. thesis statement, main idea, conclusion), entity sentence roles, grammar errors, and word usage. Miltsakaki and Kukich [2000] have explored the role of *centering theory* [Grosz et al., 1995] in locating topic shifts in student essays. Centering theory argues that the discourse in a text contains a set of textual segments, each containing discourse entities, which are then ranked by their importance. Topic shifts are generated by short-lived topics and are indicative of poor topic development. Higgins et al. [2004] have developed a system that computes similarity across text segments based on their type of discourse element and semantic similarity (LSA). A support vector machine (SVD) uses these features to capture breakdowns in coherence due to relatedness to the essay question and relatedness between discourse elements. More recently, Burstein et al. [2010, 2013a] showed how

Table 2.1

A comparison of the key features of the state-of-the art AEE systems.

AEE System	Attr. Types	Methodology	Prediction Model
PEG [Page]	Style	Statistical	multiple linear regression
PS-ME [Mason and Grove-Stephenson]		NLP	linear regression
e-rater [Burstein et al.]	Style & Content	NLP	linear regression
IntelliMetric [Schultz]			multiple mathematical models
Bookette [Rich et al.]			neural networks
OzEgrader [Fazal et al.]			
CRASE [Lottridge et al.]			machine learning
AutoScore [Shermis and Hamner]			statistical model
Lexile [Smith et al.]			Lexile measure
Ranked-based AEE [Chen et al.]			learning to rank
Multi-classifier Fusion AEE [Bin and Jian-Min]			ensemble classifiers
BETSY [Rudner and Liang]		Statistical	Bayesian networks
SEAR [Christie]			linear regression
Neural Essay Assessor [Taghipour and Ng]		Deep learning	recurrent neural networks
AES using NN [Alikaniotis et al.]			
AEG using MN [Zhao et al.]			memory networks
LightSIDE [Mayfield and Rosé]	Content	Statistical	machine learning
IEA [Foltz et al.]		LSA, NLP	
Semantic-tree-based AEE [Chali and Hasan]		LSA, tree kernel functions	cosine similarity
GLSA based AEE [Islam and Hoque]		GLSA	
Markit [Williams and Dreher]		NLP, PMT	linear regression
SAGrader [Brent et al.]	Semantic	FL, SN	rule-based expert systems
OBIE-based AEE [Gutierrez et al.]		OIE, DL	/
SAGE		OIE, NLP	random forest

the Barzilay and Lapata [2008] algorithm can be applied to the domain of student essays. In Barzilay and Lapata [2008] approach, entities (nouns and pronouns) are represented by their sentence roles and the algorithm counts all possible entity transitions between adjacent sentences in the text. By combining those entity-based features with features related to grammar errors and word usage, Burstein et al. [2013b] improve the performance of automated coherence prediction for student essays.

To best of our knowledge there are only two prior studies exploring the idea of using complex network representations for automated essay grading [Antiqueira et al., 2007; Ke et al., 2016]. In both approaches the authors used word adjacency networks to derive network-based features for essay coherence evaluation. Antiqueira et al. [2007] showed that there are strong correlations between basic structural network metrics (e.g. the average node degree, the clustering coefficient and the characteristic path length) and text quality scores assigned by human judges on a corpus of Portuguese essays written by high-school students. Ke et al. [2016] obtained similar results for a corpus of Chinese essays written by college students in a recently published study.

The above approaches mainly focus on the local coherence, while our system SAGE measures coherence as a function of semantic relatedness not only between adjacent sentences, but through the entire essay. Our proposed system measures changes between sequential essay parts from three different perspectives: semantic distance (e.g. distance between consecutive parts of an essay, maximum distance between any two parts, etc.), central spatial tendency/dispersion, and spatial autocorrelation in semantic space. Moreover, we evaluate coherence with another, i.e. network based approach. In contrast to the previous two network-based studies, our approach is based on complex networks capturing similarities between sentences.

2.3 *Detection of semantic errors in an essay*

Only two mentioned systems [Brent et al., 2010; Gutierrez et al., 2014] partially check if the statements in an essay are correct. SAGrader, developed by Brent et al. [2010], was the first AEE system that detected semantic information in an essay and upon which we based the architecture of our system. For SAGrader, the teacher first specifies the assignment prompt and desired features along with relationships among them. Using fuzzy logic, the system recognizes word combinations that can be used by students to detect desired features and relationships. Desired knowledge in the form of a semantic network is then compared with the knowledge detected in a student's essay.

The system scores the student's essay based on the similarities between observed and desired knowledge using procedural rules. Detailed feedback indicates what student did right and wrong [Brent et al., 2010].

Gutierrez et al. [2011, 2014, 2017] later proposed a system that not only detects the desired (correct) knowledge but also detects incorrect knowledge using logic reasoning [Gutierrez et al., 2014]. The system extracts statements using Open Information Extraction (OIE) and adds them to the domain ontology. The extracted tuples are in a form that is compatible with the OWL ontology. In the final step, the system determines the correctness of a statement through an ontology-based consistency checking. If the domain ontology becomes inconsistent after the extracted sentence is added into it, then the sentence is incorrect with the respect to the domain [Gutierrez et al., 2014]. Despite many efforts, this system is still not fully automated, as it requires manual inputs from the user to build a vocabulary mapping mechanism between extracted entities and vocabulary of the ontology.

In contrast to other systems, our proposed system focuses on completely automated semantic evaluation and provides semantic feedback to students. SAGE analyses text consistency by detecting entities in an essay, considering coreferences of entities, and extracting relations between entities. SAGE exploits common sense knowledge ontologies, taxonomies, and can therefore work on different domains.

In the following we overview the related work concerning the methodology we used for detecting the semantic errors in an essay.

2.3.1 Open information extraction

Information extraction (IE) is the task of automatically acquiring knowledge by transforming natural language text into structured information, such as a knowledge base [Wimalasuriya and Dou, 2010]. The main tasks of information extraction are *entity recognition* (ER), *relation extraction* (RE), and *coreference resolution* (CR). We focused on a tool for relation extraction called Open Information Extraction (OIE). Wu and Weld [2010] define the OIE system as a function that maps an unstructured document text d, to a set of triples, `{<arg1,rel,arg2>}`, where the `args` are noun phrases and `rel` is a textual fragment indicating an implicit, semantic relation between the two noun phrases. Unlike other relation extraction methods focused on a predefined set of target relations, the open information extraction paradigm is not limited to a small set of target relations known in advance, but extracts new types of relations found in

the text. The main properties of OIE systems are domain independency, reliance on unsupervised extraction methods, and scalability to large amounts of text [Gamallo, 2014].

Gamallo [2014] categorized the existing OIE systems in four groups. First he divided them into two broad categories: systems that require training data to learn a classifier and systems based on hand-crafted rules or heuristics. In addition, each former category can be divided in two subsequent types: systems that use the shallow syntactic analysis (e.g. part-of-speech tagging and/or chunking), and systems that use dependency parsing (transforming sentences into dependency trees).

For implementing semantic consistency checking in our system, we used four different OIE systems. One of the systems, Open IE [Etzioni et al., 2014], belongs to the group that needs training data and uses shallow syntax. The other three systems, ClausIE [Corro and Gemulla, 2013], CSD-IE [Bast and Haussmann, 2013], and DepOE [Gamallo et al., 2012] belong to the group that relies on rules and uses dependency parsing. We describe their use in detail in Section 3.3.

We also used a system for entity recognition (Illinois Shallow Parser [Punyakanok and Roth, 2001]) and two coreference resolution systems (Illinois Coreference Resolution [Bengtson and Roth, 2008] and Stanford Parser [Chen and Manning, 2014]) in our system for automated semantic error detection. We will further explain their application in Section 3.3.

2.3.2 *Ontology consistency*

An ontology defines a set of representational primitives with which one can model a knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application [Gruber, 2009]. A semantic network is a graphical notation for representing knowledge in patterns of interconnected nodes and arcs. To formally represent knowledge and use it in our system we use the Web Ontology Language (OWL) [Bechhofer et al., 2004]. Description logic (DL) models concepts (e.g. `Person`), roles (e.g. `isMarriedTo`) and individuals (e.g. `alice`, `bob`), and their relationships (e.g. `alice:Person`, `(alice,bob):isMarriedTo`). The fundamental modelling concept of a DL is an axiom - a logical statement that relates roles and/or concepts using conjunction, disjunction, existential and value restrictions.

Several reasoners exist for DL, one of them is HermiT [Motik et al., 2009]. The main function of a reasoner is to determine if a given knowledge base (given as an ontology) is consistent. Reasoners detect classes that are unsatisfiable, i.e. when there is a contradiction in the ontology that implies that the class cannot have any instances (OWL individuals). While OWL ontologies with unsatisfiable classes can be used, inconsistency is a severe error: most OWL reasoners cannot infer any information from an inconsistent ontology. When faced with an inconsistent ontology, they will report this and abort the classification process.

2.4 Grader effects

Due to graders' inconsistency and unreliability, it is very difficult to score essays objectively. Scores are therefore subjective and influenced by *grader* (or *rater*) *effects* which describe the influence of human factors on an assessment score. *Grader bias* and *grader reliability* are the most frequently examined grader effects when it comes to assessing essays. There is a strong empirical evidence of some graders being biased, meaning being noticeably more severe or lenient. Researchers usually use a multifaceted Rasch model to show the graders' unreliability [Congdon and McQueen, 2000; Elder et al., 2007; Myford and Wolfe, 2009].

The second most common studied grader effect is *central tendency*. Researchers provide evidence in favour of graders being biased away from the extreme scores (e.g. [Engelhard, 1994; Myford and Wolfe, 2009; Leckie and Baird, 2011]). Leckie and Baird [2011] also found no significant differences in bias or reliability between more and less experienced graders (*grader experience*).

An interesting and commonly observed grader effects are also *category/rubric preferences*. Eckes [2008] reported that graders differed significantly in their views on the importance of the various scoring criteria, resulting in several distinct types of graders. Rezaei and Lovorn [2010] report that graders were significantly influenced by mechanical characteristics of students' writing rather than the content, even when they used a scoring rubric.

The authors of the above related work verified their results using various statistical analyses. In our work we propose a novel approach for detecting different grading logics that is based on explanation methodology [Štrumbelj et al., 2009]. In the proposed approach we aim at separating the original dataset into subsets representing different graders. By modelling each grader (subset) independently, we assume that we will be

able to improve the grading accuracy, compared to modeling the entire dataset that contains mixed graders.

3

Semantic Automated Grader for Essays (SAGE)

Our aim is to develop a semantic AEE system with the motivation that the system shall (1.) improve the grading accuracy and (2.) provide an automatic feedback about the essay's semantic. Improving an existing AEE system is not an option since the majority of the systems are proprietary and LightSIDE, the only publicly available system, uses only a limited amount of mostly syntax attributes. Thus, we decided to build a new AEE system. We developed a new automated essay grading system in four phases:

1. *Automated Grader for Essays (AGE)*: The system with only linguistic and content attributes (described in Section 3.1),
2. *Automated Grader for Essays+ (AGE+)*: system AGE, augmented with additional coherence attributes (described in Section 3.2),
3. *Semantic Automated Grader for Essays (SAGE)*: system AGE+, augmented with additional consistency attributes and automatic semantic feedback (described in Section 3.3),
4. *Semantic Automated Grader for Essays- (SAGE-)*: system AGE, augmented with additional consistency attributes (but not also with the coherence attributes) and automatic semantic feedback (described in Section 3.4).

We illustrate the hierarchy of the used and newly proposed attributes in the above four AEE systems in Figure 3.1 and describe them in detail in the following sections.

3.1 Automated Grader for Essays (AGE)

We first developed a basic AEE system based on the attributes described in the literature. We named the system *Automated Grader for Essays (AGE)*. In this section, we present the used common syntax (linguistic and content) attributes that are illustrated as the right branch in Figure 3.1.

3.1.1 Syntax attributes

To implement the baseline AEE system we used 72 different attributes that were mentioned in Section 2.1. A high number of attributes measures different aspects of each essay, e.g. it has been shown that essay length significantly influences the human rater's score and we also know that the length of an essay has the highest influence on the

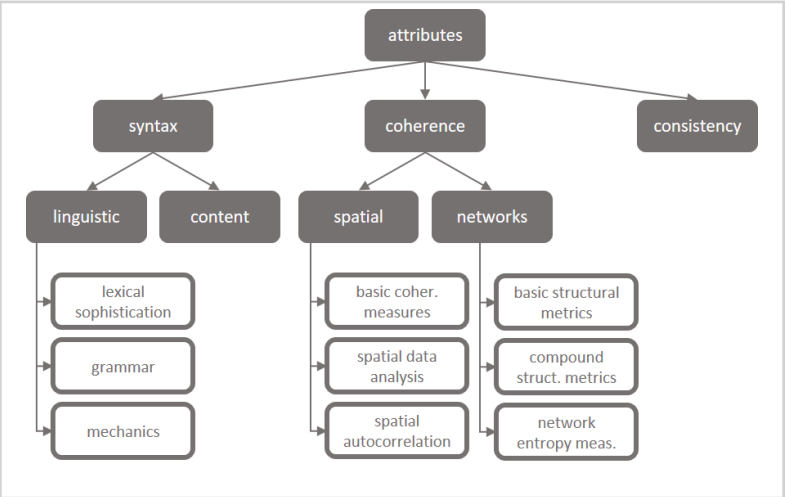

Figure 3.1

Groups of attributes used in the propsed AEE systems.

final computer score [Perelman, 2014]. Different readability measures and variations of the used words also impact the final score.

To group similar syntax attributes, we divide them into two groups:

Linguistic attributes. The linguistic attributes describe *lexical sophistication* and *grammatical* and *mechanical* aspects of the essay. These attributes are measured by counting all words, long words, different words, and number of words with different part of speech (PoS) tags. More complex attributes measure the readability level [Dubay, 2007; Smith and Jönsson, 2011], lexical diversity [Mellor, 2011], and spellchecking/capitalization/punctuation errors.

Content attributes. The second group contains the content attributes, which are based on comparing the lexical content of unseen essays with the lexical content of the graded ones using cosine similarity. To extract these attributes, we perform several comparisons of the new, ungraded essay; e.g. with the source text, with all already graded essays, with the groups of essays graded with the same grade.

3.1.2 *Presentation of the syntax attributes*

All linguistic and content attributes that we implemented in our baseline AES system are presented in Table 3.1. Overall, we extracted 72 different linguistic and content attributes. For a better presentation, we further divided linguistic attributes into three subgroups (lexical sophistication, grammar, mechanics).

3.2 *Automated Grader for Essays+ (AGE+)*

We augmented the proposed system AGE from Section 3.1 with coherence attributes and named the system Automated Grader for Essays+ (AGE+). We developed two kinds of coherence attributes that are represented with the middle branch in Figure 3.1:

1. coherence attributes, obtained in a highly dimensional semantic space (described in Section 3.2.1), and
2. coherence attributes, obtained from a sentence-similarity networks (described in Section 3.2.3).

3.2.1 *Coherence attributes obtained from a semantic space*

We base our coherence attributes on the assumption that the semantic content of a coherent essay changes gradually through its textual representation, as it has already been stated by Foltz [2007]. We start by preprocessing the essays, namely removing the numbers, punctuation, and stopwords. Then, we transform the text into lower-case letters and perform stemming. We continue with dividing essays into many sequential overlapping parts, obtained by moving a window through an essay by steps of 10 words (illustrated in Figure 3.2). Window size is defined so that it contains 25% of the average number of words per essay. For example, if the average essay length in a dataset was 280 words and the length of the essay was 320 words, we obtained 26 parts.

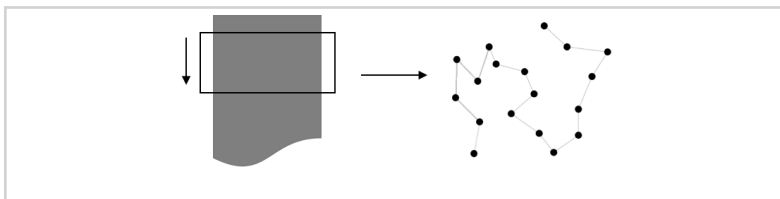

Figure 3.2

Transformation of sequential overlapping essay parts into a high dimensional semantic space [Zupanc and Bosnić, 2017a].

Table 3.1

Lexical sophistication and grammar attributes.

lexical sophistication	*grammar*
1. number of characters,	29. number of different PoS tags
2. number of words,	30. height of the tree presenting sentence structure,
3. number of long words,	31. correct verb form,
4. number of short words,	32. number of grammar errors,
5. most frequent word length,	
6. average word length,	*Number of each PoS tag*
7. number of sentences,	33. coordinating conjunction,
8. number of long sentences,	34. numeral,
9. number of short sentences,	35. determiner,
10. most frequent sentence length,	36. existential there,
11. average sentence length,	37. preposition/subordinating conjunction,
12. number of different words,	38. adjective,
13. number of stopwords,	39. comparative adjective,
Readability measures [Dubay, 2007; Smith and Jönsson, 2011]	40. superlative adjective,
	41. ordinal adjective or numeral,
	42. modal auxiliary,
14. Gunning Fox index,	43. singular or mass common noun,
15. Flesch reading ease,	44. plural common noun,
16. Flesch Kincaid grade level,	45. singular proper noun,
17. Dale-Chall readability formula,	46. plural proper noun,
18. automated readability index,	47. preposition,
19. simple measure of Gobbledygook,	48. participle,
20. LIX,	49. predeterminer,
21. word variation index,	50. genitive marker,
22. nominal ratio,	51. personal pronoun,
Lexical diversity [Mellor, 2011]	52. possessive pronoun,
	53. adverb,
23. type-token-ratio,	54. comparative adverb,
24. Guiraud's index,	55. superlative adverb,
25. Yule's K,	56. particle, "to" as preposition or infinitive marker,
26. the D estimate,	57. verb - base form,
27. hapax legomena - number of words occurring only once in a text,	58. verb - past tense,
	59. verb - gerund/present participle,
	60. verb - past participle,
28. advanced Guiraud,	61. verb - 3rd person sing. present,
	62. wh-determiner,
	63. wh-pronoun,
	64. wh-adverb.

Table 3.1

(continued) Mechanics and content attributes.

mechanics	*content*
65. number of spellchecking errors, 66. number of capitalization errors, 67. number of punctuation errors,	68. cosine similarity with source text, 69. score point level for maximum cosine similarity over all score points, 70. cosine similarity with essays that have highest score point level, 71. pattern cosine [Attali, 2011], 72. weighted sum of all cosine correlation values [Attali, 2011].

For each essay corpus (dataset) we compute the *term frequency - inverse document frequency* (TF-IDF) representation, which is a numerical statistic that reflects how important a word is to a document in a corpus. Term frequency $\mathrm{TF}(t,d)$ is computed by counting frequency of a term t in a document d. The inverse document frequency $\mathrm{IDF}(t,D)$ expresses how rare the term t is across all documents D:

$$\begin{aligned}\text{TF-IDF}(t,d,D) &= \mathrm{TF}(t,d)\cdot\mathrm{IDF}(t,D) = \\ &= \frac{|\{t\in d\}|}{|\{w\in d\}|}\cdot\log\frac{|\{d\in D\}|}{|\|\{d\in D : t\in d\}\|}.\end{aligned} \tag{3.1}$$

To compute the TF-IDF vectors of individual essay parts, we modify the computation of the TF-IDF to normalize words weights within each individual essay part with the word frequency of an entire corpus. TF-IDF vectors of essay parts represent points in high-dimensional *semantic space* that, according to our assumption, should be close to each other in coherent essays. An example of an essay divided into parts that can be visualized as points in semantic space is illustrated in Figure 3.2, in which the thin gray lines connect the points that represent the sequential parts of an essay. In Figures 3.3-3.6 we use the same example to illustrate the definition of our semantic coherence attributes, which we explain in the following subsections that outline three groups of attributes: (1) basic coherence measures, (2) spatial data analysis, (3) spatial autocorrelation.

Basic coherence measures

Basic coherence measures measure the distance between parts of the essay, which are represented as points in the semantic space. We use two variants of each attribute in this group: one computed using the Euclidean distance metric and the other computed using the cosine similarity. (In the following we will use the term "distance" to interchangeably denote any of the two.) The proposed attributes are:

- *average distance between neighbouring points* in semantic space (denoted by thin grey lines in Figure 3.3). Foltz [2007] has already shown by measuring cosine similarity between sentences in an essay that highly coherent discourses have small movements in semantic space and vice versa. We defined similar attributes that describe the average distance between these points;

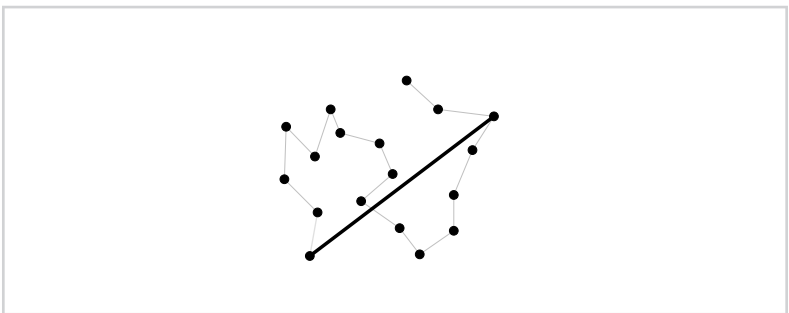

Figure 3.3

Construction of semantic coherence attributes: maximum distance between any two points [Zupanc and Bosnić, 2017a].

- *minimum* and *maximum distance* between neighbouring points and their *quotient*;
- *average distance between any two points*, which measures how well an idea persists within the essay;
- *maximum distance between any two points* measures the diameter of area that is covered with points and thus the breadth of the discussed concept in the space (illustrated in Figure 3.3);
- *Clark and Evans [1954] distance to the nearest neighbour* of each point in the semantic space for measuring spatial relationships:

$$R = \frac{\frac{\sum_{i=1}^{N} r_i}{N}}{\frac{1}{2\sqrt{N}}} = \frac{2\sqrt{N}\sum_{i=1}^{N} r_i}{N} \tag{3.2}$$

where r_i is the Euclidean distance from a given point to its nearest neighbour (see Figure 3.4) and N is the number of points. It is the measure of the degree to which the observed distribution differs from random expectation with respect to the nearest neighbour [Clark and Evans, 1954];

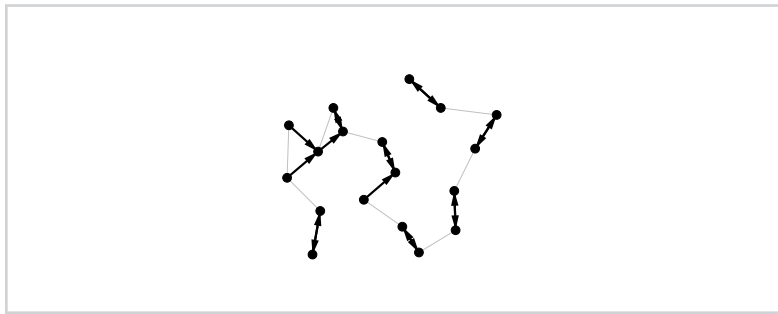

Figure 3.4

Construction of semantic coherence attributes: computation of the average distance to the nearest neighbor. Arrows indicate the nearest neighbor of each point [Zupanc and Bosnić, 2017a].

- *average distance to the nearest neighbour*, which measures how fast an idea develops across an essay (see Figure 3.4);
- *cumulative frequency distribution* G of the nearest neighbours' distances:

$$G(r) = \frac{|r \leq \bar{r}|}{N} \tag{3.3}$$

where r is the distance from a given point to its nearest neighbour (see Figure 3.4), $\bar{r}$ is the average distance to the nearest neighbour, and N is the number of points. The measure expresses the percentage of content deviations from the main idea.

Spatial data analysis

The second group of attributes describes spatial characteristics of the data and aims to extract implicit knowledge, such as spatial statistics and patterns. We adjusted a set

of descriptive spatial statistics for the use within our representation of the essays. The proposed attributes measure the central spatial tendency and the spatial dispersion, and are defined as follows:

- *average Euclidean distance between the centroid and each point*, which measures an amount of dispersion in a point-pattern (see Figure 3.5);

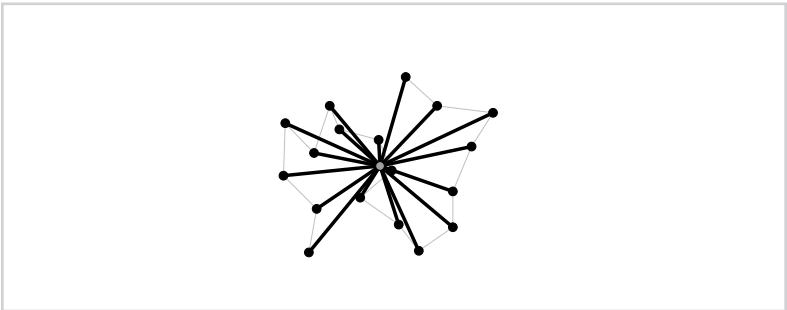

Figure 3.5

Construction of semantic coherence attributes: distance from all points to the centroid [Zupanc and Bosnić, 2017a].

- *minimal* and *maximal Euclidean distance between the centroid and each point* and *their coefficient*, which measures the biggest content deviation from the main idea;

- *standard distance* (a spatial equivalent of standard deviation) measures an amount of absolute dispersion in a point-pattern:

$$S_D = \sqrt{\frac{\sum_{k=1}^{n}\sum_{i=1}^{N}\left(D_i^k - \overline{D_c^k}\right)^2}{N}} \tag{3.4}$$

where $D_i^k, k = 1, ..., n; i = 1, ..., N$ is a k-th coordinate component of point i, $\overline{D_c^k}$ is a k-th coordinate component of a mean center, n is the number of dimensions, and N is the number of points. Similar to the standard deviation, the standard distance is also strongly influenced by extreme values. Because distances to the mean center are squared, the atypical points have a dominant impact on the magnitude of this metric, which allows detecting deviating (incoherent) essay parts;

- *relative distance*, a descriptive measure of the relative spatial dispersion. We compute it by dividing the standard distance with a measure that describes the area that is covered with points:

$$R_D = \frac{S_D}{d_{max}} \tag{3.5}$$

where d_{max} is a maximum distance of any point from the centroid. This enables direct comparison of the dispersion of different point patterns from different areas, even if the areas are of varying sizes;

- *the determinant of the distance matrix*, a measure of spatial dispersion. This allows us to measure dispersity of the content and consequently how broad the discussed topic is.

Spatial autocorrelation

Measures of spatial autocorrelation express how data tends to be clustered together in space (positive spatial autocorrelation) or dispersed (negative spatial autocorrelation). They enable us to detect global and local semantic coherence of the essays' content. If the essay exhibits positive spatial autocorrelation, this indicates that it is well structured and that the parts of the essay are well related to each other.

Typical *measures of spatial autocorrelation* are *Moran's I* [Moran, 1950], *Geary's C* [Geary, 1954], and *Getis's G* [Getis and Ord, 1992]. We adjusted these three measures so we can use them in our high-dimensional semantic space as follows:

- *Moran's I* assesses the overall clustering pattern. The original measure is intended for a 2-dimensional space, however, in this work, we adjust it to a high-dimensional semantic space by averaging it over dimensions:

$$I = \frac{N}{S} \cdot \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (D_i^k - \overline{D_c^k})(D_j^k - \overline{D_c^k})}{\sum_{i=1}^{N} (D_i^k - \overline{D_c^k})^2} \right] \tag{3.6}$$

where $D_i^k, k = 1, \ldots, n; i = 1, \ldots, N$ is a k-th coordinate component of point i, $\overline{D_c^k}$ is a k-th coordinate component of a mean center, n is the number of dimensions, N is the number of points, and S is a sum of all weights w_{ij}. Weights w_{ij} are assigned to every pair of points, with value $w_{ij} = 1$, if i and j are neighbours, and value $w_{ij} = 0$ otherwise. The range of I varies from -1 to $+1$. A positive sign of I indicates positive spatial autocorrelation and means that neighbouring points cluster together, while the opposite is true for the negative sign. Values close to zero indicate complete spatial randomness.

- *Geary's* C is inversely related to Moran's I. In this case, the interaction is not a cross-product of the deviations from the mean, but the deviations in intensities of each observation location from one another. Again, our adjusted measure is calculated in a high-dimensional semantic space and is averaged over all dimensions:

$$C = \frac{(N-1)}{2} \cdot \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (D_i^k - D_j^k)^2}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (D_i^k - \overline{D_c^k})^2} \right] \tag{3.7}$$

where $D_i^k, k = 1, \ldots, n; i = 1, \ldots, N$ is a k-th coordinate component of point i, $\overline{D_c^k}$ is a k-th coordinate component of a mean center, n is the number of dimensions, N is the number of points, and w_{ij} are point weights as described previously.

- *Gettis's* G enables us to examine point patterns at a more local scale. Gettis's G measures overall concentration or lack of concentration of all pairs of values (D_i, D_j), such that i and j are within the distance d of each other. We adjusted the measure to use it in a high-dimensional space:

$$G(d) = \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(d) D_i^k D_j^k}{\sum_{i=1}^{N} \sum_{j=1}^{N} D_i^k D_j^k} \right] \tag{3.8}$$

where $D_i^k, k = 1, \dots, n; i = 1, \dots, N$ is a k-th coordinate component of point i, n is the number of dimensions, N is the number of points, and d is the average distance between any two points in the semantic space. A weighting function $w_{ij}(d)$ is used to assign binary weights to every pair of points, where $w_{ij}(d) = 1$, if i and j are within distance d and $w_{ij}(d) = 0$, otherwise (illustrated in Figure 3.6).

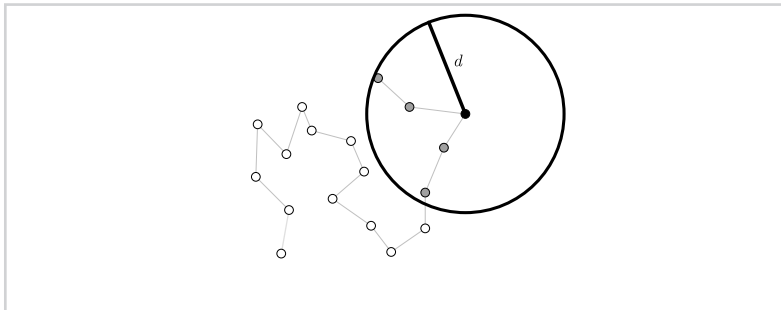

Figure 3.6

Construction of semantic coherence attributes: computation of Gettis's G using average distance d between any two points in the semantic space [Zupanc and Bosnić, 2017a].

3.2.2 *Presentation of the spatial coherence attributes*

We extracted 29 different coherence attributes from a highly dimensional semantic space and describe them in Section 3.2.1. Attributes are presented with the left part of the middle branch in Figure 3.1 and are listed in Table 3.2. Note that the attributes that were extracted twice, once using the Euclidean distance and once using the cosine similarity, are in Table 3.2 denoted with "2x".

3.2.3 *Coherence attributes obtained from sentence-similarity networks*

As a part of our collaboration with the University of Novi Sad we researched an idea of using networks to analyse coherence of student essays. The work described in this section is a result of a joint work [Zupanc et al., 2017]. We derived different structural metrics from sentence-similarity networks and used them to improve our system.

A sentence-similarity network is an undirected, weighted graph, describing similarities among sentences of an essay. Therefore, we represent each sentence as a node and use weights to represent similarity between sentences, as shown in Figure 3.7. Two nodes, representing distinct sentences A and B, are connected if their similarity is

Table 3.2

List of novel coherence attributes obtained from a higly dimensional semantic space. The attributes that are denoted with "(2x)" were computed twice, once using the Euclidean distance and once using the cosine similarity.

	basic coherence measures
1–2	average distance between neighbouring points (2x),
3–4	minimum distance between neighbouring points (2x),
5–6	maximum distance between neighbouring points (2x),
7–8	index (minimum distance/maximum distance) (2x),
9–10	average distance between any two points (2x),
11–12	maximum distance between any two points (2x),
13.	Clark's and Evans' distance to nearest neighbour,
14.	average distance to nearest neighbour,
15.	cumulative frequency distribution,
	spatial data analysis
16–17	average distance between points and centroid (2x),
18–19	minimum distance between points and centroid (2x),
20–21	maximum distance between points and centroid (2x),
22–23	index (minimum distance/maximum distance) (2x),
24.	standard distance,
25.	relative distance,
26.	determinant of distance matrix,
	spatial autocorrelation
27.	Moran's I,
28.	Geary's C,
29.	Getis's G.

higher than a given threshold w. The weight of the link connecting A and B is equal to the similarity between them.

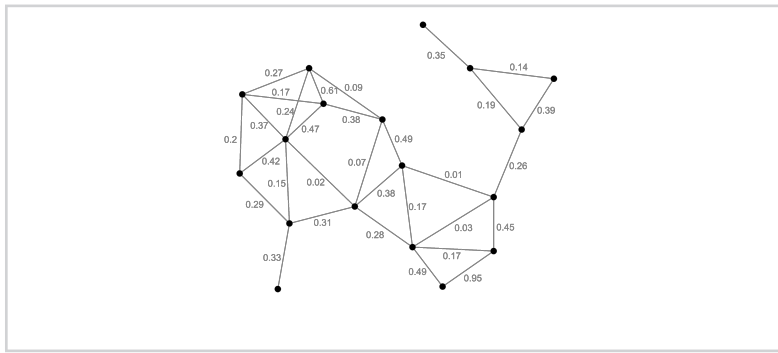

Figure 3.7

Representation of an essay with a sentence-similarity network.

To compute similarities between sentences we transform them to data points in a highly dimensional semantic space and use a distance metric to quantify their similarity. To achieve this, we compute the sentence-based TF-IDF [Robertson, 2004] representation described in Equation (3.1) and compute *cosine similarity* between their vectors. The sentence-based TF-IDF is computed by computing the document-based TF-IDF and normalizing the weights of words within each essay part with the word frequency of an entire essay. The obtained TF-IDF vectors represent points in a high-dimensional semantic space, which should, according to our assumption, be close to each other in coherent essays. We can form the final network either by connecting sentences in the decreasing order of the computed similarities until the network becomes a connected graph (i.e. a graph with a single connected component) or choose a fixed threshold. Here we decided to set the threshold equal to zero and thus obtain a relatively dense network representations of essays.

We use the obtained sentence similarity networks to compute more than 30 metrics to quantify coherence of essays. Those network-based coherence measures can be divided into the following three categories: (1) basic structural metrics, (2) compound structural metrics, and (3) network entropy measures.

Basic structural metrics

Basic structural metrics are related to characteristics of connected components, local connectivity of nodes, transitivity of links and global density and compactness of networks [Boccaletti et al., 2006]. Two nodes belong to the same connected component if they are directly or indirectly connected (i.e. there is a path connecting them). Isolated nodes (nodes that are not connected to any other nodes) also count as connected components. Connected components of a network can be determined using basic graph traversal algorithms, such as breadth-first search (BFS) or depth-first search (DFS). To quantify essay coherence we use the following five metrics related to connected components of sentence-similarity networks:

- *NCOMPS*, number of connected components;
- *LCN*, number of nodes in the largest connected component;
- *LCL*, number of links in the largest connected component;
- *LCS*, size of the largest connected component normalized by the total number of nodes;
- *ISO*, number of isolated nodes normalized by the total number of nodes.

If a sentence-similarity network contains a large number of small connected components and/or isolated nodes, then the corresponding essay contains a large number of unrelated sentences. Consequently, higher values of LCN, LCL and LCS and lower values of NCOMPS and ISO indicate a higher coherence of text.

Dense and compact sentence-similarity networks, in which nodes exhibit high local connectivity and links exhibit high transitivity, are indicators of highly coherent essays. The local connectivity of a node can be quantified by its *degree* – the number of links incident to the node. The strength of connections of a node to its neighbours in the network can be expressed by the weighted degree – the sum of weights of links incident to the node. The transitivity of links in the network is commonly quantified by the clustering coefficient which is defined as the probability that two neighbours of a randomly selected node are neighbours among themselves. The compactness of a network can be measured by the average shortest path length and the

diameter of the network – which is the longest of all shortest paths. Since sentence-similarity networks are weighted graphs, we compute both unweighted and weighted density, degree, transitivity and shortest path measures [Antoniou and Tsompa, 2008; Boccaletti et al., 2006; Costa et al., 2007]:

- *LNR*, total number of links divided by the total number of nodes;
- *DEN*, regular density which is the number of links of the network divided by the maximal possible number of links the nodes can form (denoted by M);
- *WDEN*, weighted density which is the sum of weights of all links divided by M;
- *ADEG*, average node degree;
- *AWDEG*, average weighted node degree;
- *CC*, the clustering coefficient [Watts and Strogatz, 1998];
- *WCC*, the weighted clustering coefficient;
- *ASPL*, average shortest path length considering all pairs of nodes in a largest connected component. This metric is normalized by the size of the largest connected component;
- *AWSPL*, average weighted shortest path length. Similarly as ASPL, AWSPL is computed considering all pairs of nodes in the largest connected component. The shortest weighted path between two nodes can be determined using the Dijkstra's algorithm. Also, weighted shortest paths are computed on a mirror network – the network in which the weight of the link connecting sentences a and b is equal to $1–S$, where S is the similarity between a and b;
- *DIAM*, the diameter of the network;
- *WDIAM*, the weighted diameter of the network, using the same procedure as for AWSPL when determining weighted shortest paths;
- *WIENER*, the Wiener connectivity index defined as the sum of the lengths of shortest paths for all pairs of nodes in the network;

- *WIENERW*, the weighted Wiener connectivity index defined as the sum of the weights of shortest paths for all pair of nodes in the network;

- *ZAGREB*, the second Zagreb connectivity index [Das and Trinajstić, 2011], which quantifies the degree of assortative mixing, i.e.:

$$ZAGREB = \sum_{(a,b)\in L} d(a)d(b), \tag{3.9}$$

  where L is the set of links of the network, (a, b) is the link connecting nodes a and b, and $d(x)$ denotes the degree of node x;

- *ZAGREBS*, the weighted Zagreb connectivity index where weighted node degrees are used instead of regular degrees;

- *RANDIC*, the Randić connectivity index [Yang and Lu, 2011] which quantifies the degree of disassortative mixing, i.e.:

$$RANDIC = \sum_{(a,b)\in L} \frac{1}{\sqrt{d(a)d(b)}}; \tag{3.10}$$

- *RANDICS*, the weighted Randić connectivity index where weighted node degrees are used instead of regular degrees;

- *TLS*, the total strength of links which is equal to the sum of weights of all links in the network. It was considered because sentence-similarity networks, which contain links with high weights, also indicate highly coherent texts;

- *CS*, chain strength which is equal to the sum of weight of links connecting consecutive sentences in the essay; including due to same assumption as TLS.

Compound structural metrics

Compound structural metrics are refinements of the TLS metric in which a relative importance is assigned to the network links. Namely, we can classify links of sentence-similarity networks according to the following criteria:

- *Distance of sentences in the essay* where we can distinguish between short-range and long-range links. The presence of long-range links indicates higher global

coherence of the essay, thus the weight of those links can be considered more important compared to the weights of short-range links.

- *Global centrality* where we can distinguish between bridge links connecting disjoint groups of nodes and non-bridge links. The presence of bridge links is of the utmost importance to the overall connectedness of the network, therefore they can be considered more important than non-bridge links.

- *Local centrality* where we can distinguish between local bridge links connecting loosely coupled neighbours of two directly connected nodes and local non-bridge links, which connect nodes that have a large number of common neighbours. Obviously, local bridge links are more important to local essay coherence.

According to the above mentioned criteria we propose three different adjusted TLS metrics that have the following general form:

$$WTLS = \sum_{(a,b)\in L} w(a,b) f(a,b), \tag{3.11}$$

where L is the set of links of the network, (a,b) is the link that connects nodes a and b, $w(a,b)$ is the weight of (a,b), and f is a function quantifying the relative importance of links in the network. The adjusted TLS metrics are:

- *WTLS_RD* where f is the relative distance between two sentences in the essay computed as the number of sentences between them increased by one,

- *WTLS_BS* where f is the link betweenness centrality metric. The betweenness centrality of a link e is defined as the sum of the fraction of all-pairs shortest paths that pass through e [Costa et al., 2007],

- *WTLS_CC*, for which $f = 1/CC$ where CC denotes the link clustering coefficient metric. The clustering coefficient of a link is defined as the number of common neighbours of the nodes connected by the link divided by the total number of their neighbours [Costa et al., 2007].

Network entropy measures

One of the main characteristics of complex real-world networks is a high heterogeneity of their degree distributions [Miltsakaki and Kukich, 2000]. The degree distribution of a network summarizes the local connectivity of all nodes in the network. It can be given by the probability mass function P, where $P(k)$ is equal to the probability that the degree of a randomly selected node is equal to k. The heterogeneity of the degree distribution can be quantified by the network entropy metric defined as

$$ENTR = -\sum_{k=1}^{m} P(k) \log P(k), \tag{3.12}$$

where m is the maximal node degree [Costa et al., 2007]. The minimal value of ENTR, $\text{ENTR}_{\text{min}} = 0$, is achieved whenever all nodes in the network have the same degree. Higher values of ENTR imply higher diversity of node degrees.

In addition to the entropy of the degree distribution we also compute entropies of the distributions of the following node and link metrics:

- *ENTR_BC*, the entropy of the node betweenness centrality distribution;
- *ENTR_BCL*, the entropy of the link betweenness centrality distribution;
- *ENTR_D*, the entropy of the distribution of D, where D is the relative distance between two sentences connected in the network;
- *ENTR_SD*, the entropy of the distribution of SD, where SD is the sum of relative distances between a sentence and its neighbours in the network;
- *ENTR_F*, the entropy of the node farness distribution. The farness of a node is equal to the sum of geodesic distances (the length of a shortest path) between the node and all other nodes in the network.

3.2.4 Presentation of the network coherence attributes

As a result, we extracted 32 different coherence attributes from sentence-similarity networks that are listed in Table 3.3. We joined the 29 spatial coherence attributes and 32 network coherence attributes with 72 syntactic attributes in the system AGE+.

Table 3.3

List of coherence attributes obtained from a sentence-similarity networks.

basic structural metrics
1. # connected components (NCOMPS),
2. # nodes in the largest connected component (LCN),
3. # links in the largest connected component (LCL),
4. largest connected component (LCS),
5. # isolated nodes (ISO),
6. # links divided by # nodes (LNR)
7. regular density (DEN),
8. weighted density (WDEN),
9. average node degree (ADEG),
10. average weighted node degree (AWDEG),
11. clustering coefficient (CC),
12. the weighted clustering coefficient (WCC),
13. average shortest path length (ASPL),
14. average weighted shortest path length (AWSPL),
15. diameter (DIAM),
16. weighted diameter (WDIAM),
17. Wiener connectivity index (WIENER),
18. weighted Wiener connectivity index (WIENERW),
19. second Zagreb connectivity index (ZAGREB),
20. weighted second Zagreb connectivity index (ZAGREBS),
21. Randić connectivity index (RANDIC),
22. weighted Randić connectivity index (RANDICS),
23. strength of links (TLS),
24. chain strength (CS),
compound structural metrics
25. distance of sentences in the essay (WTLS_RD),
26. global centrality (WTLS_BS),
27. local centrality (WTLS_CC),
network entropy measures
28. entropy of the node betweenness centrality distribution (ENTR_BC),
29. entropy of the link betweenness centrality distribution (ENTR_BCL),
30. entropy of the distribution of the relative distances (ENTR_D),
31. entropy of the distribution of the sum of the relative distances (ENTR_SD),
32. entropy of the node farness distribution (ENTR_F),

3.3 Semantic Automated Grader for Essays (SAGE)

Accurately reproducing the human graders is no longer the main goal of AEE systems [Bejar, 2011; Attali, 2013; Williamson et al., 2012]. It is desirable that the AEE systems can recognize certain types of errors, including syntactic errors, and offer automated feedback on correcting these errors. In addition, the systems shall also provide global feedback on content and development. The current limitation of the feedback is that its content is limited to the syntactic aspect of the essay while neglecting the semantic aspects. Exceptions are systems [Gutierrez et al., 2014; Brent et al., 2010] that include semantic evaluation of the content, but are not automatic.

When assessing student essays we can detect different kinds of errors: grammatical errors, lexical errors, semantic errors, and mechanics errors [Wu and Garza, 2014]. The AGE system already detects the grammatical, lexical, and mechanics errors with linguistic attributes (see Section 3.1). The semantic errors can be further divided into detecting the wrong word choice and detecting contradictions in the text. In this section we propose a fully automatic system that enhances AGE+: the system discovers semantic errors focusing on discovering contradictions in the text and provides a comprehensive feedback. We name the system SAGE - Semantic Automated Grader for Essays.

3.3.1 Automatic error detection system

The crucial novelty of the system SAGE is the Automatic Error Detection (AED) system. The logic of the proposed AED system is illustrated in Figure 3.8 and described in Algorithm 1. The system starts by constructing the base ontology based on common sense knowledge (everyday universal facts) and supplements it using a source text (facts in text about which the students need to write), domain knowledge (facts about a specific domain) and target knowledge (additional facts about the knowledge that the students are required to show). In parallel to constructing the base ontology, we use entity recognition, coreference resolution and open information extraction to obtain extractions from an input essay. The result of open information extraction are triples `{<arg1,rel,arg2>}` that describe relations `rel` between arguments (subjects or objects) `arg1` and `arg2`. The system afterwards proceeds by iteratively adding extractions into the base ontology and using the Hermit logical reasoner [Motik et al., 2009] to determine if an ontology is consistent after adding each extraction. If the

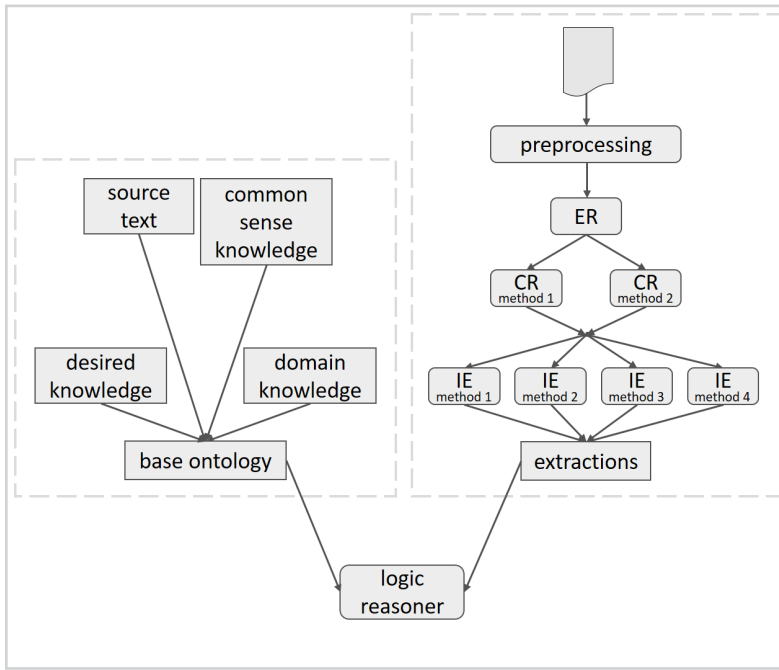

Figure 3.8

Automatic Error Detection (AED) system. The ontology-building (left) part automatically builds an ontology by combining different ontologies into a *base* ontology. The extraction (right) part extracts all possible extractions using entity recognition (ER), coreference resolution (CR) and open information extraction (IE) tools. In the final step, the AED system merges extractions, one by one, with the base ontology. If the logical reasoner determines that new extraction is inconsistent with the base ontology, it reports a semantic error [Zupanc and Bosnić, 2017a].

consistency-checking algorithm finds a contradiction in the ontology, it reports a discovered consistency error and includes it in the final feedback.

In the following subsections we describe each part of our system, illustrated in Figure 3.8, in detail.

3.3.2 *Construction of the base ontology*

The system starts building the base ontology with an ontology that contains the common sense knowledge (also referred to as an *upper ontology*). We use the Common Semantic Model (COSMO) ontology [Cassidy, 2009], which is comprised of a lattice of ontologies that serve as a set of basic logically-specified elements (classes, relations, functions, instances). The ontology is derived from elements in the public ontologies

OpenCyc[1], SUMO[2], BFO[3] and DOLCE[4]. The COSMO ontology[5] serves as a foundation ontology that has enough fundamental concept representations so that it can translate assertions from different ontologies into a common terminology and format. The COSMO ontology is in the OWL format and contains inference rules in a form of subclass and subproperty relations and restrictions [Cassidy, 2009].

We use the WordNet taxonomy [Miller, 1995] to add synonyms (gathered in *synsets*) and hypernyms to our ontology. We proceed by supplementing the COSMO ontology with the following:

1. *Source text knowledge*: Our system extracts the knowledge of the source text, upon which the essay subject is based. It processes it using steps described in Section 3.3.4. If the ontology becomes inconsistent after a new extraction is added, we detect the error and disregard the extraction.

2. *Domain knowledge*: In addition to source text knowledge, the system supplements the base ontology with domain knowledge that contains knowledge about the wider scope of an essay in a form of an ontology, including synonyms and hypernyms. For example, if students write an essay about genes and biology, we add a Gene Ontology[6].

3. *Target knowledge*: The source text and the domain knowledge represent knowledge about a specific domain. Professors or assessors can add specific desired knowledge which they explicitly expect the students to express in an essay in a form of triples. The presence of the target knowledge in an essay can have an important role when grading an essay. Detection of this knowledge can increase the accuracy of grading and improve the feedback quality.

Ambiguity is a problem inherent to language that cannot be ultimately resolved. However, we do try to disambiguate words by considering their context. When our system adds new knowledge to the ontology and it encounters ambiguity, the system

[1] http://www.opencyc.org/
[2] http://www.ontologyportal.org/
[3] http://www.ifomis.uni-saarland.de/bfo/
[4] http://www.loa-cnr.it/DOLCE.html
[5] http://www.micra.com/COSMO/
[6] http://geneontology.org/

proceeds as follows: It first compares the context (i.e. sentence) in which an entity appears in an essay with the meaning and examples of the possible instances (or classes) from the ontology. The meaning is available through the WordNet inclusion in the ontology. The instance with the most similar context use is selected. In the rare case of zero similarity, the WordNet lemma with the lowest index is selected.

3.3.3 *Processing of the ungraded essay*

Preprocessing. In the preprocessing phase the system first reads an essay and breaks it into sentences. Then it creates a duplicate of each sentence (the system needs to retain the original sentence for ER, CR, and OIE) and does several preprocessing steps: tokenization; part-of-speech tagging; finding and labelling stopwords, punctuation marks, determiners and prepositions; transformation to lower-case; and stemming.

Entity recognition. Shallow parsing is the process of identifying syntactical phrases in natural language sentences. A shallow parser identifies several kinds of phrases (chunks) that are derived from parse trees; i.e. noun phrase (NP), verb phrase (VP), prepositional phrase (PP), adverb phrase (ADVP), and clause introduced by subordinating conjunction (SBAR). These chunks provide an intermediate step to natural language understanding. Although identifying the whole parse trees can provide deeper analyses of the sentences, it is a much harder problem [Punyakanok and Roth, 2001].

Our system uses the Illinois Shallow Parser [Punyakanok and Roth, 2001] to determine chunks which we can later use for coreference resolution, searching for a suitable chunk when connecting extractions with a parts of sentences, and matching chunks with individuals, classes and relations within the ontology.

Coreference resolution. A given entity – representing a person, a location, or an organization – can be mentioned in text in multiple and even ambiguous ways. Understanding natural language and supporting intelligent access to textual information requires identifying whether different entity mentions are actually referencing the same entity [Bengtson and Roth, 2008]. The coreference resolution processes unannotated essay text and shows which mentions are coreferential.

Our system uses two different coreference resolution systems (the Illinois Coreference Resolution [Peng et al., 2015] and the Stanford Parser [Manning et al., 2014]) to detect coreferences in an essay and use them when adding extractions to the ontology.

The system combines coreferences discovered by both systems and thus increases the accuracy of discovered coreferences.

Open information extraction. After the coreference resolution, the system performs information extraction using four different systems and returns triples, as we have described in Section 2.3.1. Within the process, the duplicate extractions are removed, as well as the faulty extractions (e.g. those consisting only of a subject and a relation, while an object is missing).

After all of the previously described phases, the system starts to process each sentence sequentially and adds each extraction to the ontology by utilizing the logic reasoner.

3.3.4 Logic reasoner

After obtaining the base ontology and the extractions from the essay, we can start with discovering semantic errors. To achieve this we use Hermit, the logic reasoner [Motik et al., 2009] (described in Section 2.3.2), as shown in Algorithm 1. If the ontology is inconsistent or has an unsatisfiable concept after adding an extraction, the system concludes that there is a semantic error in the essay. The system remembers where the error occurred to later provide detailed feedback. The system then deletes the relation from the ontology and continues with the next extraction. Whenever an extraction is processed, the system first looks for both entities (predicates) in the ontology. If the ontology does not yet contain any of them, the WordNet [Miller, 1995] taxonomy is used to find synonyms or coreferenced entities within the ontology. If the ontology still does not contain any synonyms or coreferences, the system looks for hypernyms of the entity and creates a subclass (i.e. creating a triplet with *subclassOf* relation). If all described attempts fail, the last alternative is to create a new class or individual in the ontology (see Section 2.3.2). When both entities are included in the ontology, the system first checks if the specific relation is not yet a part of the ontology and adds it accordingly.

Algorithm 1

Pseudocode of the Automated Error Detection (AED) system.

```
function MAIN(common_sense_ontology, domain_knowledge, target_knowledge,
source_text, ungraded_essay)
    function EXTRACT(text, ontology)
        Preprocessing
        Entity recognition
        Coreference resolution
        Open information extraction
        for sentence in text do
            for relation in sentenceRelations do
                Add to ontology
                HermiT check
                print errors
            end for
        end for
        return (ontology, errors)
    end function

    ontology ← common_sense_ontology
    (ontology, __) ← EXTRACT(source_text, ontology)
    ontology ← ontology + domain_knowledge
    ontology ← ontology + target_knowledge
    (__, errors) ← EXTRACT(ungraded_essay, ontology)
    return errors
end function
```

3.3.5 Presentation of the consistency attributes

Based on detections of semantic errors in an essay that were detected by the logic reasoner (as explained in the previous section) we implemented three error attributes that are illustrated as the right branch in Figure 3.1:

1. *number of unsatisfiable cases* when adding classes and individuals in the ontology,
2. *number of inconsistency errors* after adding a triple to ontology,
3. *total number of consistency errors* (sum of the first two attributes).

In the remaining sections we proceed to evaluate the benefits and performance of the proposed attributes.

We extracted 3 different semantic attributes that are listed in Table 3.4. We joined these attributes with 133 attributes from AGE+ into a system SAGE.

Table 3.4

Consistency attributes.

semantic
1. number of unsatisfiable cases
2. number of inconsistency errors,
3. total number of consistency errors,

3.3.6 Providing automated feedback

One of the main advantages of the system SAGE is that it provides comprehensive and informative feedback about syntactic and semantic errors. When it detects an error, it reports a pair of conflicting ontological relations to the student. The automated error detection is based on the logic reasoner, which we evaluate in the following. We also provide some examples of a semantic feedback from the AED system.

Performance of the automated error detection system

To preliminarily evaluate the proposed automated error detection system, we constructed an artificial dataset consisting of 50 sentences describing a girl named Lisa.

We manually labelled sentences as correct and incorrect to denote the ground truth, having 36 correct and 14 incorrect sentences. As an input to the automated error detection system we used only a common sense ontology and were aiming to measure how effectively the system detects incorrect sentences.

We measured sensitivity and specificity of our system, where the sensitivity expresses the proportion of incorrect sentences that are correctly identified as incorrect, and the specificity measures the proportion of correct sentences that are correctly identified as correct. By running the experiment, we obtained 100% specificity and 71.4% sensitivity. The 100% specificity was expected, since the system treats each sentence as correct unless it detects an error in the sentence. The sensitivity shows that there is still a room for improving successful detection of incorrect sentences, and we know where to focus on.

The pipeline of the proposed AED system consists of preprocessing, information extraction, and a logic reasoner, where information extraction is further performed in three steps: entity recognition, coreference resolution, and relation extraction. Note, that our system can be only as good as the weakest tool in our pipeline. Even though assessing the quality of each single step is not our objective, we have to be aware of the quality of the NLP tools. The most questionable are the IE tools. The system used for entity recognition (Illinois Shallow Parser [Punyakanok and Roth, 2001]) reports the F1-score of approximately 0.92, and the two coreference resolution systems (Illinois Coreference Resolution [Bengtson and Roth, 2008] and Stanford Parser [Chen and Manning, 2014]) report F1-scores of 0.80 and 0.60, respectively. For higher consistency, we used different approaches for the same task, however, especially the performance of the coreference resolution is modest. We argue here, that our aim was not to improve existing NLP tools, but to achieve the best possible results with the existing ones. Therefore, we aim our system to detect as many errors as possible and we aim to improve the accuracy of our system when novel approaches for information extraction tasks will be proposed. We can furthermore also improve the detection of ambiguous sentences and sentences that require reasoning by including many different relations in the ontology. This subject shall be the focus of further research.

Examples of the provided feedback

In the following we first provide several examples obtained on the artificial dataset described above. We then proceed with the examples obtained on the real world datasets.

Figure 3.9 displays a simple example of a student who was writing about a girl Lisa. When he wrote that Lisa is a boy, our system detected an error and reported it in the form of feedback. The system discovered the error because the classes `girl` and `boy` are subclasses of `female` and `male` classes, respectively, that are disjoint hypernyms, but are yet taken into account by SAGE.

```
Extraction ['Lisa', 'is', 'a boy'] is inconsistent with
a relation ['Lisa', 'is', 'girl'] in the base ontology.
```

Figure 3.9

Error detection in case of disjoint hypernyms [Zupanc and Bosnić, 2017a].

In Figure 3.10 we can see an example that first uses coreference resolution to detect that she and Lisa are synonyms. Furthermore, the relation `born` is a function in our ontology, i.e. each class can have at most one relation of type `born`. When another relation is added for the same class, system detects an error.

```
Extraction ['she', 'was born', 'in London'] is inconsistent with
a relation ['Lisa', 'born', 'Paris'] in the base ontology.
```

Figure 3.10

Error detection by considering coreferences and unique relations.

Figure 3.11 provides a third example in which the student first wrote that Lisa does not like sports and later that she likes tennis. In the ontology, `tennis` is a subclass of `sport` and relations `like` and `not like` are disjoint. Coreference resolution detected that she and Lisa are synonyms, so the system recognized an error. Notice that an ontology and the system allow that a class has more relations of the same type with different classes (e.g. `like` or `isA` relation) as long as these classes are not disjoint (e.g. Lisa can be a girl and a student).

```
Extraction ['She', 'likes', 'tennis'] is inconsistent with
a relation ['Lisa', 'not like', 'sport'] in the base ontology.
```

Figure 3.11

Error detection by considering synonyms and coreferences [Zupanc and Bosnić, 2017a].

The fourth example (in Figure 3.12) represents a more complex example as a combination of three sentences. First two sentences: "Lisa likes slow sports and doesn't like quick sports." and "Tennis is a quick sport." do not initiate an error. When a student

writes a sentence "Lisa likes tennis." the system returns an error. As mentioned before in the ontology, `tennis` is a subclass of `sport` and relations `like` and `not like` are disjoint. Likewise, the classes `slow` and `quick` are disjoint hypernyms, based on which the system is able to detect an error and return the feedback.

```
Extraction ['Lisa', 'likes', 'tennis'] is inconsistent with
relations ['Lisa', 'like', 'slow sport'] and
['Tennis', 'is', 'quick sport'] in the base ontology.
```

Figure 3.12

Error detection of disjointness of hypernyms [Zupanc and Bosnić, 2017a].

To test how the proposed system works on the real world data, we ran it on the source-based essays written by 13- and 15- year-old students (for more details about the data see Secton 4.1.1). We manually checked the output of several essays to report on the AED system's performance.

The first source was a short story titled *Rough Road Ahead: Do Not Exceed Posted Speed Limit* by *Joe Kurmaskie*. Students had to write a response that explains how the features of the setting affect the cyclist. We present a part of an essay where the student was writing about the cyclist and wrote:

> The setting was hot and dry, which affected the cyclist greatly. She didn't have enough space to carry a lot of water.

The coreference resolution detected that *the cyclist* and *she* are referring to the same person, consequently meaning that the cyclist is a woman. The base ontology - among other knowledge - includes data from the source text, where we extracted the fact that the cyclist is a man named Joe Kurmaskie. The error is shown in Figure 3.13.

```
Extraction ['Joe Kurmaskie', 'is', 'a man'] is inconsistent with
a relation ['the cyclist', 'is', 'a woman'] in the base ontology.
```

Figure 3.13

Error detection using coreference resolution.

Another feature of the proposed system is that it checks for the facts or objects that have to be discussed in an essay. We input this information through the desired knowledge part of the ontology. In one of the source-based essays students were writing a response about the story *The Mooring Mast* by *Marcia Amidon Lüsted* where they had to describe the obstacles faced by the builders of the Empire State Building. The

most important was to mention the dirigibles. The following short response did not exhaustively address the problem:

> The Empire state building was facing of with Chrysler building that was being constructed. Chrysler building had a trick up his sleeve by constructing 185 foot spire inside the building and then shocked the public. Bring it to a hieght of @DATE1 feet, 46 feet taller than the originally announced hieght of the Empire State building. Soon to be the tallest building. The empire state building was destined to never fulfill it's purpose.

The system reports on the several missing pieces of information in the text as seen in Figure 3.14.

```
In your essay you did not write about 'dirigibles'.
To improve your essay you can include the following facts:
['steel frame', 'has to be', 'strengthened']
['dirigibles', 'cannot fly', 'low']
['dirigibles', 'dock', 'in open landing fields']
```

Figure 3.14

Detection of the missing information based on the desired knowledge part of the ontology.

However, because the pipeline of the proposed system is long and several tools achieve only moderate accuracy, we expected to detect false positive and false negative examples. The first presented error occurs due to the mistake in the coreference resolution step. The student is writing about a girl and her mother from the story *Winter Hibiscus* by *Minfong Ho*. In the sentence

> Saeng's mother understood how she was feeling, and she was not dissapointed in her for failing her driving test either.

the student wrote that the mother was not disappointed. But the coreference resolution refers to *she* in the first and *she* in the second part as the same entity even though the first one is referring to the girl and the second one is referring to her mother. Consequently, the system adds to the ontology that the girl is not disappointed. The ontology previously included the fact from the source text that the girl was disappointed, thus the system returns an error (see Figure 3.15).

```
Extraction ['she', 'was not', 'disappointed'] is inconsistent with
a relation ['Saeng', 'is', 'disappointed'] in the base ontology.
```

Figure 3.15

An example of a false positive error in an essay due to the error in the coreference resolution.

In the last example we provide a semantic mistake that remains undiscovered. A student wrote the following sentence:

> She says that she much rather do gardening then school work.

There was no such statement in the text, but because it did not contradict any fact from the ontology, the semantic mistake remained undiscovered.

3.4 Semantic Automated Grader for Essays- (SAGE-)

Semantic Automated Grader for Essays- (SAGE-) focuses on evaluation of the consistency of the facts written in an essay and thus includes only syntax and consistency attributes described in Sections 3.1 and 3.3. We included this system for evaluation purposes only to research the contributions of the consistency attributes in detail. Since the syntax and consistency attributes are already described in previous sections, we omit repeating them here.

4

Comparison of AGE, AGE+, SAGE-, and SAGE against the state-of-the-art

The four proposed systems in Section 3 represent four different aspects of grading:

- scoring an essay without understanding the content (AGE),
- evaluating the content through the coherence (AGE+),
- prioritizing the semantic of an essay with provided feedback to a student (SAGE), and
- prioritizing the consistency of the facts written in an essay (SAGE-).

To evaluate the systems we first extracted the proposed attributes from the text. For extracting the syntax attributes, we helped ourselves by using the *Natural Language Toolkit* (NLTK) [Bird et al., 2009] for natural language processing in Python and a spellchecking library *PyEnchant*[1]. In this chapter we describe the methodological details of the grading models and how we compared and evaluated the systems.

4.1 Evaluation

4.1.1 Essay datasets

We performed the experiments on datasets that were provided within the Automated Essay Scoring competition on the Kaggle website[2]. The datasets contain student essays for eight different prompts (essay discussion questions). The anonymized students were from the USA and were drawn from three different grade levels: 7, 8, and 10 (aged 12, 13, and 15, respectively). Four datasets included essays of traditional writing genres (persuasive, expository, narrative) and the other four were *source based* (i.e. the students had to discuss questions referring to a previously read source document). Each training set was pre-scored by at least two human expert graders. Since Dataset 2 was scored using two different criteria, it appears as two separate datasets 2a (scored with an emphasis on writing skills) and 2b (scored with an emphasis on language skills) in the tables with the results.

The authors of the datasets already divided them into fixed training and test sets. We used training sets during the attribute development phase: for syntax and spatial coherence attributes we trained on dataset 1, for network coherence attributes we

[1] https://pythonhosted.org/pyenchant/

[2] Access to data can be requested through the Kaggle website http://www.kaggle.com/c/asap-aes/data or ASAP website http://www.scoreright.org/

trained on dataset 8, and for consistency attributes we trained on dataset 3. During the testing phase, we evaluated the performance also on other datasets that were not included in the development process above. We used the same training and test sets as authors of the datasets to build scoring models and measure prediction accuracy, respectively. The characteristics of the used datasets are shown in Table 4.1.

Table 4.1

Properties of essay datasets divided into training (first part) and test set (second part).

Characteristic	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
Type of essay	persuasive	persuasive	source-based	source-based	source-based	source-based	expository	narrative
Grade	8	10	10	10	8	10	7	10
# essays	1,783	1,800	1,726	1,771	1,805	1,800	1,569	723
Mean # words	366.40	381.19	108.69	94.39	122.29	153.64	171.28	622.13
SD of # words	120.40	156.44	53.3	51.68	57.37	55.92	85.2	197.08
Range of grades	2-12	1-6 1-4	0-3	0-3	0-4	0-4	0-24	0-60
Mean grade	8.53	3.42 3.33	1.85	1.43	2.41	2.72	19.98	37.23
# essays	589	600	568	586	601	600	441	233
Mean # words	368.96	378.4	113.24	98.7	127.17	152.28	173.48	639.05
SD of # words	117.99	156.82	56.0	53.84	57.59	52.81	84.52	190.13
Range of grades	2-12	1-6 1-4	0-3	0-3	0-4	0-4	0-24	0-60
Mean grade	8.62	3.41 3.32	1.9	1.51	2.51	2.75	20.13	36.67

SD = standard deviation; # = number of

In the following we present two examples of written essays that are part of DS1. Essays from DS1 were written by 13-year-old students where they had to write a letter to a local newspaper and state their opinion on the effects that computers have on people. Note that the authors of the datasets automatically anonymized names in essays to prevent inference about the participating states. The first essay is an example of a well written essay (resolved score 12):

> Dear @ORGANIZATION1, @CAPS1 has been brought to my attention that some people feel that computers are bad for us. Some people say that they are a distraction to our physical and mental health. Although I can see how some people would think this, I believe that computers are a good benifit to all society. I believe this because computers can help people learn, stay intach with friends or family that live far away, and stay orginized. Sometimes people are on the computer, learning and they don't even know @CAPS1. Simply by visiting the @ORGANIZATION2 homepage, you automaticly see the news feeds of things happening around

> the world. Other times people go online diliberatly to learn. If someone is thinking about going to @LOCATION1 then they would probably go on the internet to learn about @CAPS1. Simply by searching equadore many choices will pop up you climate, sesonal weather, hotel options, and other facts. But thats not the only way people are learning on the internet. Now, many college students have the option of taking their lessons online. This is because some students like calm quietness or own house the distractions of sitting in class. Friends could be a big distraction in class, but how can you stay intouch with your friends if they moved away? I remember in second grade my bestfriend, @LOCATION2, move away. I was so sad. I badey ever talked to her, but then one day our parents set us up on a vidio chat! I felt like I was right their with her! This was great, and I though about how many people could use this to talk to relatives or friends. Another great way to stay intouch into friends and family is through e-mail. By writing a message and sending @CAPS1 can make staying in touch so easey, and your personal wants can chat and emails are a easey thing to send world wide. So many people love to type on a keyboard as well, but so many different papers that you type could be lost. I, for me, hate clutter, and I have so many school binders for papers to be lost in. This is why I take great advantage of typing my paper every chance I get. My computer keeps me orginiced because I could never loose my work. File save, is an idiot proof way to keep all your files in a safe place. Then all you have to do is press print to get a hard copy. I am sure that many people love using their computer for the same reason. Also, I myself am a much faster typer than I am writer so my work is a lot needey on the computer. As you can see their are plenty of reasons why using a computer is goof for our society you can learn, stay intouch with friends and family, and stay orginiced. Many people, could agree with me. Don't you?

The second essay is an example of a poorly written essay (resolved score 5):

> Computers don't have any affect on kids we just love going on cause we use it for help and this persuade the readers of the local newspaper cause we need to be able to communicate also do writing essays and doing social studies or science homework my ideas are let us go computers cause were not bothering u can just leave us alone and let us do what you need to do cause what computers are what give us information for we have to do and were to do wat we gotta do and u people can just leave us alone cause arent addicting to me or anyone and if we were it still would it matter cause a computers a computer u dont punish it because just punish us from the computer punish us because of it cause its the computer fault it can be addicting cause the computer is device that gives us wat we need and the information we also the computer does favors for us the computer is a amazing thing.

4.1.2 Evaluation measures

For evaluation of the prediction models performance we use two widely used performance measures:

- the *exact agreement* measure, which is defined as the percentage of essays that were graded equally by both graders.

- the *quadratic weighted Kappa*, which is an error metric that measures the degree of agreement between two graders and is analogous to the correlation coefficient. This metric typically ranges from 0 (expected agreement between random scores) to 1 (complete agreement between graders). In case that there is a lesser agreement between the graders than expected by chance, this metric can have values below 0. Assuming that a set of essays E has S different possible scores, $1, 2, \dots, S$, and that each essay receives scores from two different graders (e.g. human/computer), the metric is calculated as follows:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{4.1}$$

where w are weights, O is a matrix of observed scores and E is a matrix of expected scores. The matrix of weights w_{ij} is an S-by-S matrix that is calculated based on the difference between graders' scores, such that

$$w_{i,j} = \frac{(i-j)^2}{(S-1)^2}. \tag{4.2}$$

The matrix of *observed* scores O, which is an S-by-S histogram (agreement) matrix, is constructed over the essay scores, such that $O_{i,j}$ corresponds to the number of essays that received a score i by grader A and a score j by grader B; analogously, E is an S-by-S histogram matrix of *expected* scores:

$$E_{i,j} = \frac{H_{Ai} \cdot H_{Bj}}{N} \tag{4.3}$$

where $H_{Ai}, i = 1, \dots, S$ denotes the number of essays that grader A scored with score i, and N is a number of gradings or essays. E is normalized with N such that E and O have the same sum [Zupanc and Bosnić, 2015].

To compare the significance of the difference between two quadratic weighted Kappas, we used the Wilcoxon signed-rank test. This is a non-parametric statistical test that is used when comparing repeated measurements on a single sample to assess whether their population mean ranks differ. The test assumes that data are paired, come from the same population and are not necessarily normally distributed [Kanji, 2006].

4.1.3 Prediction model

Our first experiments to predict the final score included the following classifiers with the described key properties:

- linear regression: using QR decomposition, without regularization and parameter fitting; using *lm* from the *stats*[3] package in R;
- regression trees: with constructive induction in the inner nodes; node splitting criteria: *RReliefFexpRank*, splits are binary; in regression tree leaves we are using linear reduced models (as in M5); using *CoreModel* from the *CORElearn*[4] package in R;
- feed-forward neural networks: single-hidden-layer neural network; 6 units in the hidden layer; weight decay = 0.1; (case) weights set to 1; maximum number of iterations 1000; using the *nnet*[5] package in R;
- random forest: 100 trees; sampling of cases is done with replacement; the number of attributes randomly sampled as candidates at each split is $\frac{|attributes|}{3}$; using the *randomForest*[6] package in R;
- extremely randomized trees: a model similar to random forest, but it uses the same data to train all trees in a set and chooses splitting nodes randomly among variables; the ensemble contained 100 trees; the number of attributes tried at each node was $\frac{|attributes|}{3}$; the number of random cuts for each (randomly chosen) attribute was 1 (default), which corresponds to the official ExtraTrees method; cutting thresholds are uniformly sampled; using the *extraTrees*[7] package in R.

Table 4.2 shows the results of the Kappa metric for each classifier using all the attributes from the AGE system. We performed the evaluation using 10-fold cross-validation on the training sets. Since the results showed that the random forest and extremely randomized trees [Geurts et al., 2006] achieved the highest performance, we decided to use them as essay grade predictors in further evaluation. Both models, using random

[3] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
[4] https://cran.r-project.org/web/packages/CORElearn/CORElearn.pdf
[5] https://cran.r-project.org/web/packages/nnet/nnet.pdf
[6] http://cran.r-project.org/web/packages/randomForest/randomForest.pdf
[7] http://cran.r-project.org/web/packages/extraTrees/extraTrees.pdf

Table 4.2

Values of the quadratic weighted Kappa for different regression models: linear regression (LR), regression trees (RT), neural network (NN), random forest (RF), and extremely randomized trees (ERT).

Model	DS1	DS2a	DS2b	DS3	DS4	DS5	DS6
LR	0.8359	0.7232	0.5175	0.6535	0.7090	0.7900	0.7663
RT	0.8070	0.6943	0.4885	0.6620	0.7113	0.7828	0.7184
NN	0.8247	0.6964	0.4883	0.6328	0.6877	0.7776	0.7428
RF	0.8447	0.7389	0.5386	0.6591	0.7174	0.7949	0.7636
ERT	0.8434	0.7439	0.5384	0.6554	0.7148	0.7967	0.7670

Model	DS7	DS8	average
LR	0.7781	0.7785	*0.7280*
RT	0.7323	0.7289	*0.7028*
NN	0.7601	0.7247	*0.7039*
RF	0.7888	0.7738	*0.7356*
ERT	0.7882	0.7807	*0.7365*

forest and extremely randomized trees, continued to achieve similar results in all the following experiments, thus we report only the results for the random forest model.

4.2 Results

To analyse the potential benefits of the proposed attributes, we first evaluated their relevance and contribution to predictive accuracy. We proceed by comparing predictive accuracies of four different versions of our AEE system and continue by comparing the best system to other state-of-the-art systems.

4.2.1 Evaluation of the implemented attributes

As described in Section 3, we implemented 136 existing and novel attributes (72 linguistic and content attributes, 29 + 32 coherence attributes, and 3 consistency attributes). To improve model interpretability, achieve shorter training times and enhance generalization by reducing overfitting, we performed attribute selection to detect redundant and irrelevant attributes. Attribute selection was performed using the forward attribute selection approach. Starting with an empty set of attributes, an attribute that improves the model performance the most (measured by the quadratic weighted Kappa measure) was included into the set in each iterative step. The procedure was

terminated when there were no more attributes that improved the model performance. The model performance was measured using the internal 10-fold cross-validation on the training set, and the best attribute in each step was selected according to the highest average Kappa value among all folds.

The ranks of the 50 most relevant attributes, averaged across all data sets, are shown in Table 4.3 in the decreasing order of the average rank. From the ranking we can see that the number of words and number of different words influence the final grade the most, as well as the score point level that uses cosine similarity between already graded essays and a new essay. We can observe that some of the proposed coherence attributes rank among the top 10 attributes: Geary's C (7th), local centrality (8th), and Clark's and Evans' distance to nearest neighbour (9th). Moreover, the results show that coherence attributes present 38% of the top 50 attributes. The most highly ranked proposed content attribute – number of inconsistency errors – is in the 46th place.

We further evaluated the influence of attributes on the final score, with the intention to interpret the meaning of their values by plotting the trend dependencies. In the following we provide three representative examples from different attribute groups (syntax, coherence, and consistency groups from Figure 3.1).

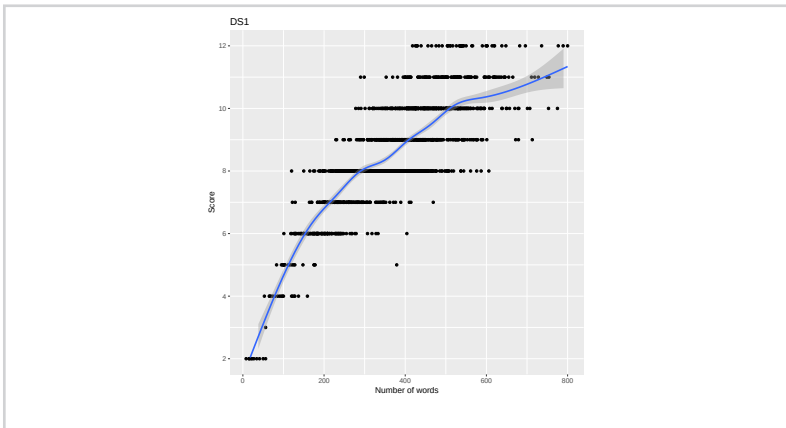

Figure 4.1

The scatter plot of the essay's score in dependency on the syntax attribute *number of words* with the smooth local regression line using DS1.

Figure 4.1 illustrates the dependency of the length of the essay (measured in number

Table 4.3

Average ranks of 50 most relevant attributes (1-25) within all 136 attributes across all 9 datasets. The ranks were obtained using the forward attribute selection. Group abbreviations stand for: syntax-linguistic (S-L), syntax-content (S-C), coherence-spatial (C-S), coherence-networks (C-N), and consistency (C).

attribute	*group*	*average rank*
1. number of words	S-L	16.44
2. number of different words	S-L	23.33
3. score point level for max cos. sim. over all score points	S-C	27.33
4. number of sentences	S-L	32.11
5. number of tokens	S-L	33.78
6. number of spellchecking errors	S-L	36.78
7. Geary's *C*	C-S	38.33
8. local centrality (WTLS_CC)	C-N	38.52
9. Clark's and Evans' distance to nearest neighbour	S-L	38.67
10. pattern cosine	S-C	39.11
11. type-token-ratio	S-L	39.89
12. number of genitive markers	S-L	41.22
13. number of characters	S-L	41.33
14. cosine similarity with source text	S-C	42.00
15. entropy of the link betw. centr. distr. (ENTR_BCL)	C-N	41.27
16. index (minimum distance/maximum distance) (Euclid)	C-S	42.89
17. average word length	S-L	43.22
18. number of verbs - past tense	S-L	43.89
19. average distance between neighbouring points (Euclid)	C-S	44.22
20. number of long sentences	S-L	44.89
21. min distance between points and centroid (cos)	C-S	47.11
22. number of predeterminers	S-L	50.44
23. number of particles	S-L	50.67
24. Getis's *G*	C-S	50.78
25. relative distance	C-S	51.11

Table 4.3

(continued) Average ranks of 50 most relevant attributes (25-50) within all 136 attributes across all 9 datasets. The ranks were obtained using the forward attribute selection. Group abbreviations stand for: syntax-linguistic (S-L), syntax-content (S-C), coherence-spatial (C-S), coherence-networks (C-N), and consistency (C).

attribute	*group*	*average rank*
26. standard distance	C-S	51.67
27. simple measure of Gobbledygook	S-L	53.11
28. weighted sum of all cosine correlation values	S-C	53.33
29. average weighted node degree (AWDEG)	C-N	53.52
30. most frequent word length	S-L	53.56
31. number of superlative adverbs	S-L	54.00
32. happax legomena	S-L	54.00
33. number of possessive pronouns	S-L	54.11
34. cumulative frequency distribution	C-S	54.22
35. number of adverbs	S-L	54.89
36. min distance between points and centroid (Euclid)	C-S	55.44
37. min distance between neighbouring points (cos)	C-S	56.56
38. max distance between any two points (cos)	C-S	56.67
39. strength of links (TLS)	C-N	56.72
40. number of superlative adjectives	S-L	56.78
41. max distance between neighbouring points (cos)	C-S	56.78
42. entropy of the node farness distribution (ENTR_F)	C-N	56.94
43. max distance between neighbouring points (Euclid)	C-S	57.00
44. number of verbs - base form	S-L	57.44
45. LIX	S-L	58.67
46. number of inconsistency errors	C	58.75
47. Yule's *K*	S-L	59.11
48. weighted 2nd Zagreb connectivity index (ZAGREBS)	C-N	59.26
49. number of participles	S-L	59.33
50. number of wh-determiners	S-L	59.56

of words) on the final score on DS1. We observed the same trend on all datasets. As it was also shown in Perelman [2014] this indicates that the higher graded essays a have higher number of words.

Coherence attributes show similar trends on different domains (i.e. datasets). However, we extracted different values for coherent essays on different domains using the same attributes. Hence, we are not able to define a scale for each attribute that would define how coherent an essay is. Figure 4.2 shows that different datasets (DS1 and DS6) have a similar trend of the Moran's I values and confirms our assumption that a coherent essay implies high positive autocorrelation, meaning that neighbouring parts tend to cluster together. In Section 4.1.1 we provided examples of a good and a bad essay. Note that the first one is a coherent essay and achieves a Moran's I score of 0.27, while the second example is a less coherent essay and achieves a Moran's I score of 0.03. It is evident for a reader that the flow of information in the second essay is disorganized and is thus hard to follow.

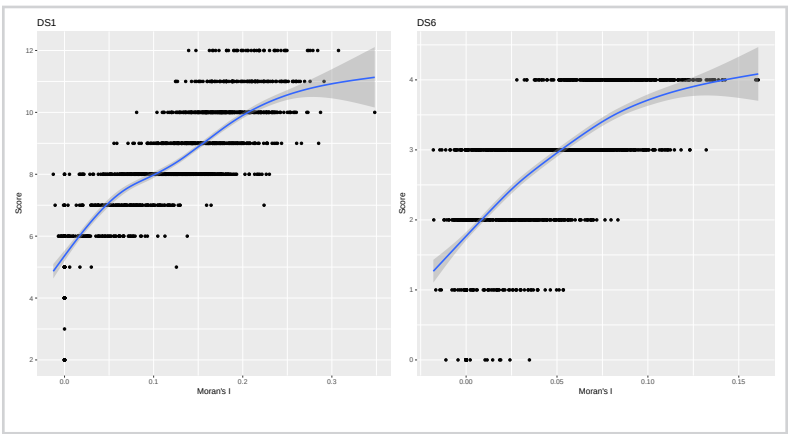

Figure 4.2

The scatter plots of the essay's score in dependency on the coherence attribute *Moran's I* with the smooth local regression line using DS1 and DS6.

Using consistency attributes we plotted the normalized total number of consistency errors (see Figure 4.3) on DS4. As expected, the trend mostly shows less mistakes for essays with higher scores and more mistakes for essays with lower scores. The peak in the beginning appears due to two reasons: (1) there are quite some short essays that do not include many semantic errors but are of low quality because of other criteria; and (2) the dataset includes a higher number of essays graded with scores 1 and 2 and

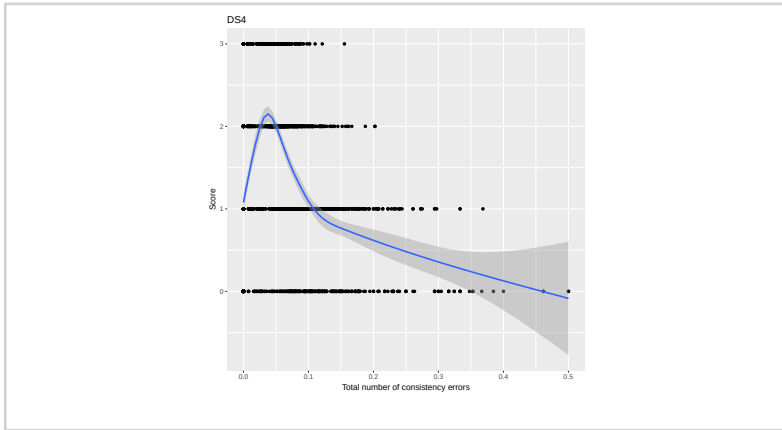

Figure 4.3

The scatter plot of the essay's score in dependency on the number of consistency errors with the smooth local regression line using DS4.

a lower number of essays scored with score 3.

Dataset 8, in addition to the final score, provides 6 rubric scores describing ideas and content, organization, voice, word choice, sentence fluency, and convention for each essay. Since the organization rubric also describes the coherence, we decided to further investigate how well our proposed coherence attributes predict the organization rubric score. In this experiment, we prepared datasets with different sets of attributes: (a) spatial coherence attributes only (29), (b) network coherence attributes only (32), (b) syntax (linguistic and content) attributes only (72), and (c) syntax and coherence attributes (133). We used random forest without attribute selection to build the prediction models. Table 4.4 shows the results of all models using the quadratic weighted Kappa and Table 4.5 shows p-values calculated between the results of all the models. Based on the high influence of number of words on the final grade, we expected high prediction accuracy already by using the set of syntax attributes only. Nevertheless, by adding the set of coherence attributes to the set of syntax attributes, the accuracy additionally increased. We can also see that both sets of coherence attributes alone also achieved relatively high prediction accuracy, which enabled us to conclude in favour of their benefit.

We additionally calculated the Spearman coefficients between the coherence attributes

Table 4.4

Comparison of the prediction accuracies using quadratic weighted Kappa for the organization rubric, which also measures coherence.

attributes	quadratic weighted Kappa
spatial coherence attributes	0.6040
network coherence attributes	0.5813
syntax attributes	0.6928
syntax and coherence attributes	0.7025

Table 4.5

Comparison of the significance of the difference between prediction accuracies for different models for the organization rubric using Wilcoxon signed-rank test.

attributes	network coherence attributes	syntax attributes	syntax and coherence attributes
spatial coherence attributes	0.0201	0.0019	0.0016
network coherence attributes		0.0020	0.0009
syntax attributes			0.0433

and the organization rubric score. Getis's G, Moran's I, and weighted Wiener connectivity index (WIENERW) achieved the highest absolute correlations with 0.5947, 0.5752, and 0.5627, respectively (p-values < 0.001). Overall 42 of 61 coherence attributes correlate with the organization rubric score with p-value smaller than 0.05.

To further research the correlations between different attributes we calculated Spearman coefficients between all attributes for all datasets. We expect the computed correlation to reveal the dependencies between attributes, as well as the relation between the attribute value and the final score. Figure 4.4 shows the heatmap with correlation coefficients averaged over all datasets. The attributes are arranged in the same order as they are represented in Tables 3.1, 3.2, 3.3, and 3.4 from left to right and from the bottom up on the x and y axis, respectively. Syntax attributes are followed first by spatial and network coherence attributes and at the end by consistency attributes. The last attribute is the resolved score, which in the first row and in the last column of the heatmap illustrates the correlations of the attributes with the human score. We can

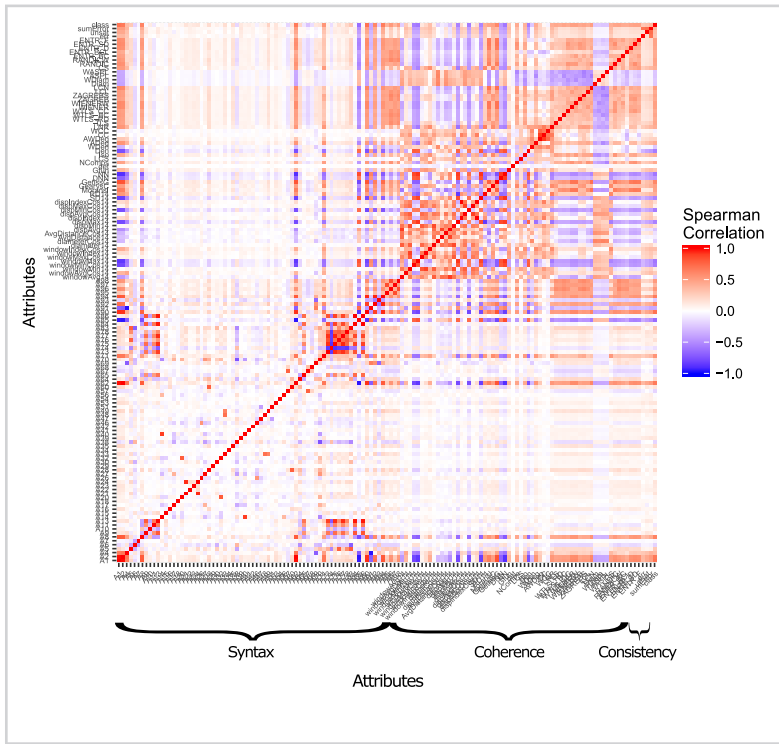

Figure 4.4

The heatmap representing the average correlations between attributes over all datasets.

observe that coherence attributes indicate higher correlation among themselves than with the other attributes. An interesting observation indicates that a group of network coherence attributes negatively correlates with others; those are average shortest path length (ASPL), average weighted shortest path length (AWSPL), weighted diameter (WDiam), and diameter (Diam). A smaller square with higher correlations among the syntax attributes represents correlations between readability measures (Gunning Fox index, Flesch reading ease, Flesch Kincaid grade level, Dale-Chall readability formula, automated readability index, and a simple measure of Gobbledygook).

To further illustrate the relations between attributes we also visualized the average correlation matrix using multidimensional scaling (MDS) [Cox and Cox, 2008] in Figure 4.5. MDS projects our 137-dimensional data (including the final score)

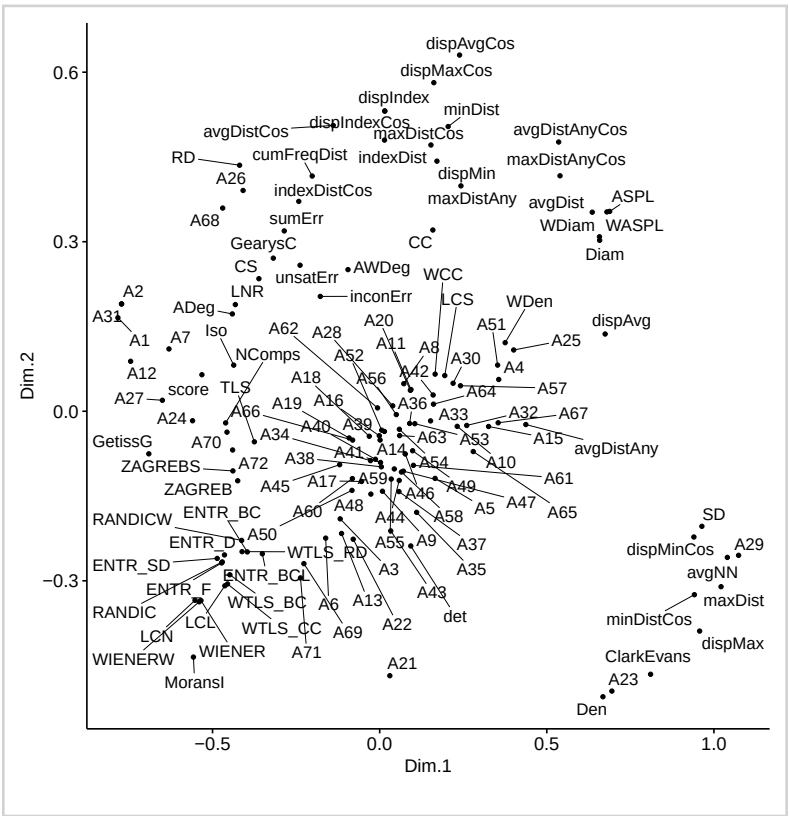

Figure 4.5

The visualization of the average correlation matrix using multidimensional scaling.

to a 2-dimensional space such that similar objects in the 137-dimensional space are close together on the 2-dimensional plot. Each attribute is presented with a dot and the name, however because several attributes appear very close to each other in a 2-dimensional space, the names of some attributes are remotely written and connected with the associated dot with a line. For higher readability, the syntax attributes are denoted with their sequential number from Table 3.1. We can observe the formation of several groups including coherence attributes, which corresponds to the earlier observation that several coherence attributes strongly (positively or negatively) correlate

to one another.

When plotting the average correlations between attributes over all datasets in Figure 4.4, we noticed differences between the strength of correlations between the same attributes on the different datasets. Thus, we calculated the correlation matrices for each dataset and performed a pairwise subtraction between matrices of each dataset. We detected small differences between pairwise coefficients in correlation matrices of datasets that contain the same type of essays (e.g. persuasive, source-based, expository, narrative) and noticeable differences between pairwise coefficients in correlation matrices of datasets with different essay types. Figure 4.6 illustrates two representative examples of matrix differences: the left one shows the differences between datasets of the same type (including persuasive essays – DS1 and DS2a), and the right example shows the differences between datasets of different types (the source-based essays and the narrative essays – DS4 and DS8, respectively). We can conclude that the correlations are domain-dependent. We ascribe the reasons for smaller correlation differences between datasets that contain the same type of essays mainly to the fact that human graders evaluate the same type of essays with the same grading instructions, meaning that for each different type of essays they prioritize different aspects of essay quality.

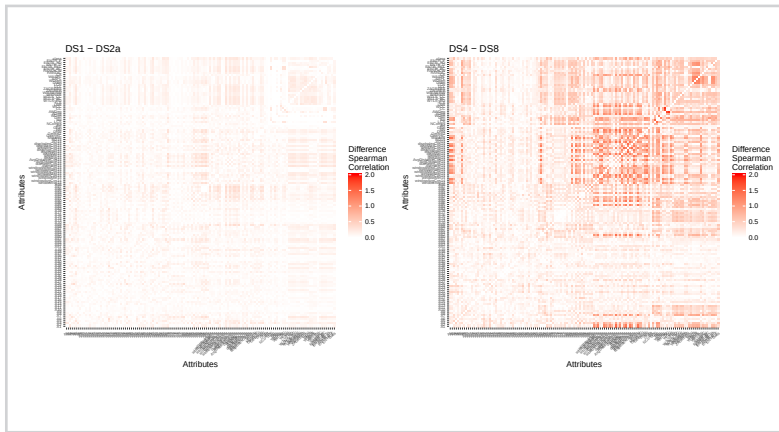

Figure 4.6

The correlation coefficient differences heatmaps between datasets of essays of the same type (left) and datasets of essays with different types (right).

4.2.2 *Accuracy of the semantic-based AEE system*

In the following experiments, we compared four versions of our system to evaluate if semantic attributes yield to better model performance: AGE, AGE+, SAGE-, and SAGE. Since the systems SAGE- and SAGE require source-based essays to build an ontology for the logic reasoner, we were able to evaluate it only on datasets that include source-based essays (such datasets are 3, 4, 5, and 6).

Table 4.6 shows the quadratic weighted Kappas and exact agreement for AGE and AGE+. We calculated the p-values between both approaches on the same dataset by running the same two models 10 times (without setting the seed). The results show that the prediction accuracy significantly (p-value$<$ 0.05) improves on 8 out of 9 datasets when the coherence attributes are used in the system. The comparison of the average results in the rightmost column of Table 4.6 shows that there is also a significant difference between both systems over all datasets (p-value $<$ 0.01).

Table 4.6

Comparison of the system AGE (syntactic attributes only) and the system AGE+ (with additional coherence attributes) using the quadratic weighted Kappa (1st row) and exact agreement (2nd row), p-values are computed for Kappas and star (★) indicates significant difference ($p < 0.05$).

System		DS1	DS2a	DS2b	DS3	DS4
AGE	QW Kappa	0.9045	0.7473	0.6619	0.8096	0.8040
	Exact agg.	0.7224	0.7716	0.7379	0.7886	0.7237
AGE+	QW Kappa	0.9251	0.7924	0.6714	0.8272	0.8109
	Exact agg.	0.7507	0.8057	0.7481	0.8036	0.7375
p-value		<0.001★	<0.001★	0.0416★	0.0116★	0.0398★

★ p-value<0.05

System		DS5	DS6	DS7	DS8	average
AGE	QW Kappa	0.8701	0.7736	0.8760	0.7851	*0.8036*
	Exact agg.	0.7805	0.7314	0.2607	0.1577	*0.6305*
AGE+	QW Kappa	0.8729	0.7817	0.8814	0.8050	*0.8187*
	Exact agg.	0.7847	0.7400	0.2627	0.2219	*0.6505*
p-value		0.0205★	<0.001★	0.0201★	0.0570	*0.0039*★

★ p-value<0.05

Table 4.7 shows the quadratic weighted Kappas and exact agreement for four systems: AGE, AGE+, SAGE- and SAGE on four source-based datasets. We calculated

Table 4.7

Comparison of the systems AGE (syntax attributes only), AGE+ (syntax and coherence attributes), SAGE- (syntax and consistency attributes) and SAGE (syntax, coherence and consistency attributes) on source-based datasets using quadratic weighted Kappa (1st row) and exact agreement (2nd row), p-values are computed for Kappas and star (★) indicates significant difference ($p < 0.05$). The red color indicates p-values where varying component is only the set of consistency attributes. Some results are copied from Table 4.6 for easier comparison.

System		DS3	DS4	DS5	DS6	average
AGE	QW Kappa	0.8096	0.8040	0.8701	0.7736	0.8143
	Exact agg.	0.7886	0.7237	0.7805	0.7314	0.7561
AGE+	QW Kappa	0.8272	0.8109	0.8729	0.7817	0.8232
	Exact agg.	0.8036	0.7375	0.7847	0.7400	0.7665
SAGE-	QW Kappa	0.8246	0.8104	0.8719	0.7796	0.8216
	Exact agg.	0.8017	0.7304	0.7813	0.7385	0.7629
SAGE	QW Kappa	0.8340	0.8120	0.8791	0.7880	0.8283
	Exact agg.	0.8100	0.7302	0.7962	0.7353	0.7679
p-value AGE – AGE+		0.0116★	0.0398★	0.0205★	<0.001★	0.0013★
p-value AGE – SAGE-		0.0337★	0.0527	0.0637	0.0172★	0.0183★
p-value AGE - SAGE		0.0086★	0.0349★	0.0071★	<0.001★	<0.001★
p-value AGE+ – SAGE-		0.1563	0.2469	0.2258	0.0758	0.0852
p-value AGE+ – SAGE		0.0549	0.1312	0.0174★	0.0073★	0.0125★
p-value SAGE- – SAGE		0.0442★	0.0789	0.0114★	0.0012★	0.0063★

★ p-value<0.05

the p-values between approaches on the same dataset by running the same models 10 times. We also used 10 Kappas for each model to calculate the p-values over all datasets. Since we already analysed the influence of coherence attributes in Table 4.6, our aim was to determine whether the consistency attributes contribute to the higher prediction accuracy. Thus, we compared the system AGE to the system SAGE- and the system AGE+ to the system SAGE with the only varying component therefore being the consistency attributes. Hence, the p-values evaluating those two comparisons in Table 4.7 are coloured red. The results show that using consistency attributes leads to higher Kappa values on all four observed datasets for the both pairs of compared systems (AGE vs. SAGE- and AGE+ vs. SAGE). Furthermore, the improvements were significant on two out of four datasets (DS3 and DS6 for the AGE versus SAGE- comparison and DS5 and DS6 for the AGE+ versus SAGE comparison) and also on the average for both comparisons (the rightmost column). Note also that the difference

between AGE+ and SAGE- is not significant on any of the datasets, indicating not enough statistical evidence to conclude whether only coherence attributes or consistency attributes in addition to syntax attributes (AGE) contribute to greater predictive performance.

We repeated the experiment from Table 4.7 using 10-fold cross-validation instead of using the fixed training and test sets. Table 4.8 reports on the results and shows the quadratic weighted Kappas and variance for four systems: AGE, AGE+, SAGE- and SAGE on four source-based datasets. We calculated the p-values between approaches on the same dataset using all 10 Kappas obtained from 10-fold cross-validation. We also used 10 Kappas for each model to calculate the p-values over all datasets. Again, we

Table 4.8

Comparison of the systems AGE (syntax attributes only), AGE+ (syntax and coherence attributes), SAGE- (syntax and consistency attributes) and SAGE (syntax, coherence and consistency attributes) on source-based datasets using 10-fold cross-validation. Quadratic weighted Kappa (1st row) and variance (2nd row) are reported, p-values are computed for Kappas and star (★) indicates significant difference ($p < 0.05$). The red color indicates p-values where varying component is only the set of consistency attributes.

System		DS3	DS4	DS5	DS6	average
AGE	QW Kappa	0.7909	0.7873	0.8692	0.7573	0.8012
	Variance	0.0012	0.0013	0.0003	0.0010	/
AGE+	QW Kappa	0.7965	0.8008	0.8767	0.7916	0.8164
	Variance	0.0012	0.0009	0.0002	0.0007	/
SAGE-	QW Kappa	0.7929	0.7914	0.8738	0.7598	0.8045
	Variance	0.0008	0.0007	0.0003	0.0010	/
SAGE	QW Kappa	0.8010	0.8072	0.8805	0.7982	0.8217
	Variance	0.0015	0.0007	0.0002	0.0009	/
p-value AGE - AGE+		0.1162	0.0157★	0.0436★	<0.001★	<0.001★
p-value AGE - SAGE-		0.3758	0.1869	0.0422★	0.3130	0.0718
p-value AGE - SAGE		0.0019★	<0.001★	0.0048★	<0.001★	<0.001★
p-value AGE+ - SAGE-		0.6921	0.0372★	0.0507	<0.001★	0.0055★
p-value AGE+ - SAGE		0.2573	0.1704	0.0123★	0.0015★	0.0259★
p-value SAGE- - SAGE		0.0489★	<0.001★	0.0158★	<0.001★	<0.001★

★ p-value<0.05

focused on the influence of the consistency attributes, thus we compared the system AGE to the system SAGE- and the system AGE+ to the system SAGE. The consistency attributes again induced higher Kappa values on all four observed datasets for

both pairs of compared systems. However, the improvements were significant only on one out of four datasets (DS5) for the AGE versus SAGE- comparison and on two out of four datasets (DS5 and DS6) for the AGE+ versus SAGE comparison. On the average (the rightmost column) the SAGE system showed a significant difference in comparison to AGE+ and SAGE- does not significantly improve the prediction accuracy compared to the system AGE. Furthermore, the results show a significant difference between the AGE+ and SAGE- systems on two out of four datasets (DS4 and DS6) and also on the average (the rightmost column), in contrary to finding in Table 4.7, indicating better contribution of coherence attributes to the predictive performance.

To verify how the proposed systems compare on real-world data, we provide an example of an essay where AGE+ performs better than AGE, and SAGE performs better than AGE+. We score the following essay with all three systems and obtain three different scores:

> This story is a heart-warming tale of how Family and community can thoroughly transform one's life. For example, the author briefly mentions the fact that his parents moved to make a better life in @LOCATION1 than the one that would've been possible in Cuba. Also, Narciso tells of the support that neighbours gave him recived, regardless of race. For example, it says that the author's house always had "open doors" to those who needed a place to stay.People often think of happiness as some thing relatively hard to achieve, but this memoir makes it clear that happiness can just be hosting a dinner for family, or letting struggling friends stay for a while. Just like in "A strange Old @CAPS1", the story shows how fruitful one's life can be by doing morally good things for people, no matter who it is.The distinct mood that someone would obtain from reading this is a particular kind of hapiness-not an excited kind like when someone wins the lottery, but the warm glowing kind when someone helps out; the satisfied kind when someone knows they've made someone else happy.

AGE predicts the score 1 based on only syntax attributes where it detects several misspellings and the length that is shorter than other well-scored essays. AGE+ predicts the score 2 taking into account syntax attributes and the coherence attributes that reveal the good coherence of the essay. SAGE predicts score 3 taking into account no detected semantic errors. The human resolved score is 3 on the scale $0 - 4$.

4.2.3 Comparison with the state-of-the-art AEE systems

We also compared the proposed system SAGE with the state-of-the-art systems that were used in a previous study [Shermis and Hamner, 2013] at the end of 2012: PEG, e-rater, IntelliMetric, CRASE, LightSIDE, AutoScore, IEA, Bookette, Lexile Writing Analyzer, with a ranked-based approach [Chen et al., 2012] and with results obtained

by researchers participating in the before mentioned Automated Essay Scoring competition on the Kaggle website. The eight commercial systems among the above listed systems capture over 97% of the current automated scoring market in the USA [Shermis and Hamner, 2013].

Tables 4.9 and 4.10 show the results that were calculated between the automated and human scores (resolved score of more human graders). Since not all the systems are available for public experimenting, their results were obtained from the papers [Shermis and Hamner, 2013] and [Chen et al., 2012], and from the Kaggle website[8]. Results reported in [Shermis and Hamner, 2013] and [Chen et al., 2012] include Kappa values for every data set and are reported in Table 4.9 together with results of SAGE. The evaluated systems are sorted in descending order of the average Kappa value, which is shown in the rightmost column of Table 4.9. Since dataset 2 has scores in two different domains (see Section 4.1.1), each transformed Kappa is weighted by 0.5. The Wilcoxon non-parametric test was used to compute p-values that express the significance of differences between each evaluated system and SAGE. We can see that our system achieves significantly better results on 5 out of 9 datasets (DS1, DS2a, DS3, DS5, DS7). On the remaining four datasets, accuracy of SAGE was insignificantly different from the accuracy of the best performing system, while still significantly better compared with some of the systems. On the average (the rightmost column), SAGE achieved significantly better results than 9 out of 10 other systems.

We also compared SAGE with results obtained from the leader board of Automated Essay Scoring competition. In Table 4.10 we ranked 8 commercial systems, 8 leading systems from the competition, LightSide [Mayfield and Rosé, 2013], ranked-based system [Chen et al., 2012] and SAGE. The results are reported in the form of the average Kappas over all datasets, since the accuracy of 8 leading systems on the Kaggle website is reported like that.

[8] http://www.kaggle.com/c/asap-aes/data

Table 4.9

Comparison of the proposed semantic grading system SAGE with other state-of-the-art systems. The table shows quadratic weighted Kappas, achieved on different datasets. Significantly different values ($p < 0.05$) are marked with a ★.

System	DS1	DS2a	DS2b	DS3	DS4	DS5
SAGE	*0.93*	*0.79*	0.67	*0.83*	0.81	0.89
PEG	0.82★	0.72★	*0.70*	0.75★	*0.82*	0.83★
e-rater	0.82★	0.74★	0.69	0.72★	0.80	0.81★
IntelliMetric	0.78★	0.70★	0.68	0.73★	0.79	0.83★
CRASE	0.76★	0.72★	0.69	0.73★	0.76★	0.78★
LightSIDE	0.79★	0.70★	0.63	0.74★	0.81	0.81★
ranked-based	0.81★	0.68★	0.68	0.67★	0.73★	0.80★
AutoScore	0.78★	0.68★	0.66	0.72★	0.75★	0.82★
IEA	0.79★	0.70★	0.65	0.65★	0.74★	0.80★
Bookette	0.70★	0.68★	0.63	0.69★	0.76★	0.80★
Lexile	0.66★	0.62★	0.55★	0.65★	0.67★	0.64★

★ p-value<0.05

System	DS6	DS7	DS8	average
SAGE	0.79	*0.88*	*0.81*	*0.83*
PEG	*0.81*	0.84★	0.73	0.79
e-rater	0.75	0.81★	0.70★	0.77★
IntelliMetric	0.76	0.81★	0.68★	0.76★
CRASE	0.78	0.80★	0.68★	0.75★
LightSIDE	0.76	0.77★	0.65★	0.75★
ranked-based	0.72★	0.77★	0.71★	0.74★
AutoScore	0.76	0.67★	0.69★	0.73★
IEA	0.75	0.77★	0.69★	0.73★
Bookette	0.64★	0.74★	0.60★	0.70★
Lexile	0.65★	0.58★	0.63★	0.63★

★ p-value<0.05

Table 4.10

Accuracy comparison of various systems from the literature and results from the Kaggle competition.

System	Avg. acc.	rank
SAGE	0.8325	1
Sollers & Gxav*	0.8014	2
SirGuessalot & PlanetThanet & Stefan*	0.7986	3
VikP & jman*	0.7978	4
Efimov+Berengueres*	0.7956	5
@ORGANIZATION*	0.7947	6
PEG [Page, 1994]	0.7888	7
Martin*	0.7857	8
cs224u*	0.7828	9
jackpot (Jason)*	0.7826	10
e-rater [Burstein et al., 2013a]	0.7656	11
IntelliMetric [Schultz, 2013]	0.7588	12
CRASE [Lottridge et al., 2013]	0.7494	13
LightSIDE [Mayfield and Rosé, 2013]	0.7494	14
Ranked-based [Chen et al., 2012]	0.7363	15
AutoScore [Shermis and Hamner, 2013]	0.7325	16
IEA [Foltz et al., 2013]	0.7344	17
Bookette [Rich et al., 2013]	0.6981	18
Lexile [Smith et al., 2014]	0.6331	19

* Results were obtained from the leader board of AES competition on Kaggle website[5].

5

Automated grouping of similar graders

We have stated already in Section 1.2 that human grading is inconsistent and unreliable. In Section 2.4 we provide some overview of the research proving that scores are subjective and influenced by grader effects, i.e. scores are affected by factors such as bias (strictness, leniency) and (un)reliability (non-systematic error) of the grader. Systematic and non-systematic human errors introduce subjective variance into scores and therefore impact their validity [Lottridge et al., 2013].

In this chapter we propose a novel approach for separation of the original dataset that contains scores given by several different graders into smaller subsets that group essays scored by the same grader. To detect different graders solely by their given grades we use the explanation methodology [Štrumbelj et al., 2009], which enables us to detect different dependencies (grading logic) between essays' attributes and its score. Further, we build an ensemble of models on the detected subsets and aim to improve the prediction accuracy in comparison to a model built on the initial joint dataset.

5.1 Explanation methodology

Štrumbelj et al. [2009]; Štrumbelj and Kononenko [2014] introduced a method (Interactions-based Method for Explanation, IME) for explaining decisions of an arbitrary regressor (classifier) on a level of each individual example. The method decomposes the model's prediction value (class) for an instance into the contributions of the attributes' values. The method is independent of the used classification algorithm and considers interactions and redundancies between attributes.

The method provides explanation in term of attribute - value contributions. The computed contributions reflect the attribute's influence on the final decision of the explained model. The contribution sign indicates if the individual attribute value affected the predicted value in a positive or a negative way. That means, positive contributions yield to a higher final score and negative contributions conversely yield to a lower score. The output explanation is a vector of contributions that represents the inner knowledge of the model.

Figure 5.1 illustrates an explanation of an individual regression prediction using the IME method. From here on we will use *nomogram* as an alternative notion for a bar chart from Figure 5.1 that represents attributes' contributions. As we can see from the text above and below the figure, the predicted score for the particular essay using the random forest model was 3.65 and the true score was 4. The bars in the chart represent the contributions of example's attribute values that are returned by the

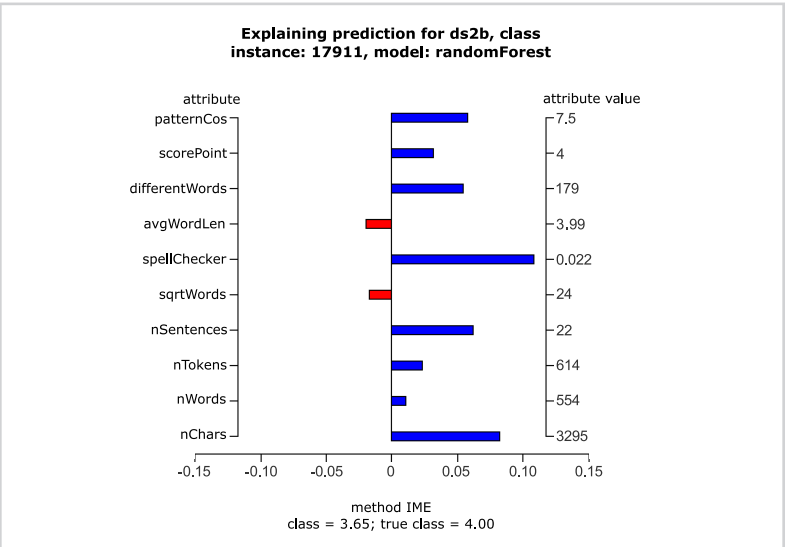

Figure 5.1

IME-based explanation of an individual example. The visualization is explaining the influence of AEE attributes on the final score (ranging from 0 to 3).

explanation algorithm. The blue coloured bars indicate the positive influence of the attribute value (higher predicted final score) on the final score and the red coloured bars indicate the negative influence of the attribute value (lower predicted final score). The attributes that strongly contribute to the higher score are: number of spell-checking errors, number of characters, weighted score of the most similar graded essays based on cosine similarity (patternCos), number of different words, and number of sentences. Several other attributes slightly improve the predicted score: score for which the max cosine correlation was obtained (scorePoint), number of tokens, and number of words. Two attributes marginally negatively affect the final score: average word length and square root of the number of words.

Since we assume that each grader has his/her own subjective grading logic and criteria, we use the presented methodology to detect grading patterns within a dataset and separate graders into smaller subsets.

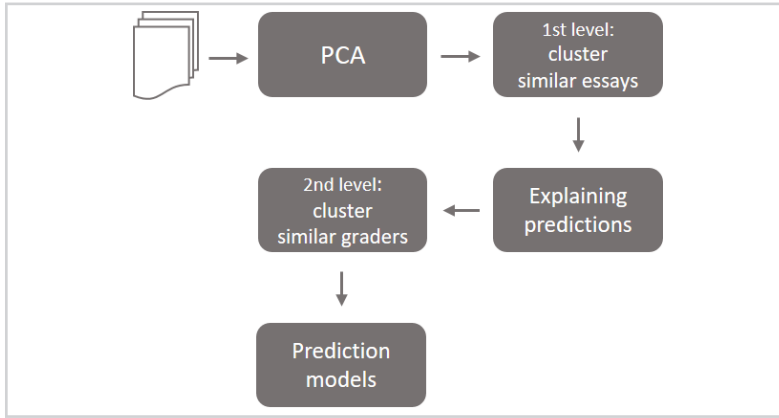

Figure 5.2

An overview of the proposed approach for dividing a DS into subsets that represent different graders. We use PCA and a two-level clustering approach. On the first and second level of clustering we cluster together similar graders inside the essay clusters. The obtained clusters serve as training sets for building the prediction models.

5.2 Detection of similar graders

In this section, we describe the methodology for dividing a dataset into subsets that represent similar graders. In the proposed methodology we lean on the underlying assumption that we can detect *different graders* when they score *similar essays* with *different scores or score explanations*.

The steps of the proposed approach are illustrated in Figure 5.2 and summarized as follows:

1. Since the essay datasets feature a high number of attributes that can impact the size of the problem space, training time and generalization/overfitting, we begin with PCA (principal component analysis) as a dimensionality reduction approach. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated attributes into a set of values of linearly uncorrelated attributes, called principal components. The number of principal components is less than or equal than the number of original attributes. This transformation is defined in such a way that the first principal component has the largest possible variance (i.e., accounts for as much of the variance in the data as possible), and each following component has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors therefore form an uncorrelated orthogonal

basis set [Abdi and Williams, 2010].

2. In the second step we perform the first-level clustering in which we group *similar essays*. We perform clustering on the essays' principal components that are computed from their attributes. Here we apply different k-means clustering approaches, as suggested by Hartigan and Wong [1979], Lloyd [1957], and Macqueen [1967].

3. In the third step we calculate the *explanations of essay grade predictions*, as described in Section 5.1. The explanations reveal the dependency between attributes and the predicted grade, i.e., grading logic and grader's criteria for each essay.

4. We proceed by performing the *second-level* clustering within each of the first-level clusters. In each cluster of similar essays, we cluster together essays according to their grading explanations. This allows us to group essays into groups that represent consistent grading criteria and therefore hopefully reflect different graders.

5. In the last step, we build prediction models for each obtained second-level cluster. To later grade a yet unseen essay the above procedure allows us to map it to the most similar first-level cluster and predict a score using different second-level cluster models.

The second-level clustering of similar essays in Step 4 groups essays according to similarity of their grading logic with the aim to detect different grading patterns. Note, that the discovered clusters in this group may either represent individual different graders or group similar graders into the same cluster. Since our main goal is neither discovery of the number of graders nor the analysis of their grading patterns, but rather improvement of grading prediction accuracy, this does not represent an obstacle.

To classify an unseen essay, we compute its prediction explanation. Based on the explanation, we find the closest existing second-level cluster and predict using the model from that cluster (or create an ensemble of multiple learners represented by multiple nearest clusters and predict the final grade using them).

In our experiments we employ two variants of the above approach. They differ in whether we use PCA only as the attribute space transformation or as an attribute selection approach for the subsequent clustering. Both variants use the same remaining workflow as shown in Fig. 5.2. We describe both variants in detail in the next subsections.

Variant 1: PCA as the attribute space transformation

The first variant transforms the attribute space and performs the level-one clustering on the transformed space. The approach is comprised of the following 5 steps, adapted from Fig. 5.2:

1. Only a number `n.components` of the most important PCA components are selected.
2. Level-one clustering is performed on the transformed attribute space using the selected components. The parameter `n.clust.l1` defines the number of required target clusters on this level.
3. We use attribute selection to lower the dimensionality of the original attribute space by selecting `n.import.attr` attributes with the highest average rank using ReliefF [Kononenko, 1994], Gini index [Gini, 1912], and information gain [Mitchell, 1997]. Afterwards, we calculate the explanations of predictions on the resulting attribute space.
4. Level-two clustering is then performed on prediction explanations. The parameter `n.clust.l2` defines the number of clusters on the second level.
5. We learn prediction models for each cluster.

Variant 2: PCA as the attribute selection approach

The second approach utilizes PCA in a different manner, as follows:

1. The PCA on the input attributes is used to calculate the impact of each original attribute on the first 50 principal components. Among the attributes that had the highest sum of contributions, `n.infl.attr` attributes are selected.
   From here further, the remaining steps follow implicitly as in Fig. 5.2.

2. Level-one clustering is performed on the resulting attribute space with lowered dimensionality.

3. We calculate explanations of predictions.

4. Level-two clustering is performed on prediction explanations.

5. We learn prediction models for each cluster.

To summarize, the main difference between both variants lies in Step 1, which also reflects in Step 3, where an additional feature selection method is used to reduce the problem space.

5.3 Experimental environment and evaluation

We perform the evaluation of the proposed approach in two steps. First, we analyse the clustering quality to determine if the approach truly detects different graders. Since the grader IDs are not known for real datasets, we produce the artificial datasets in which the true grader ID is known (but hidden from the learning algorithm). In the second step, we select the best performing approach from the first step and evaluate it on two real datasets: one dataset contains *single* individual rater scores for each essay, and the other contains *resolved scores* from multiple graders.

5.3.1 Datasets

We use real datasets (DS) and artificial datasets (ADS) that we derive from the real datasets.

Real datasets. We use the real-world datasets described in Section 4.1.1 that provide a variety of different essay types as well as datasets with single and resolved scores. All datasets provide a resolved score from at least two human graders' scores, but do not provide the graders IDs. Additionally, each essay in a dataset can be graded by different two graders.

Artificial datasets. To create a controlled environment with known grader IDs, we created artificial datasets by grading essays from the real datasets using the LightSIDE [Mayfield and Rosé, 2013] (described in Section 2.1) AES system. We used two different group of attributes: one using 2-grams and the second one using 3-grams on the lemmatized essays. We used random forest and SVM, respectively, to

build the models. We randomly assigned a score from one of these two systems to each essay, simulating a dataset that contains mixed scores from two different graders. Each dataset contained a balanced number of scores from both grading systems.

To form the datasets for supervised learning, we extracted 136 attributes that are described in Chapter 3.

5.3.2 Evaluation metrics

Clustering quality

We evaluate similarity between clusters with an internal and an external clustering validation measure:

- The *Dunn Index* [Dunn, 1974] is an internal clustering evaluation measure, which relies only on the information in the data. It measures the ratio of the smallest distance between observations from different clusters to the maximum distance between observations in the same cluster. It is defined as:

$$DI = \frac{min_{1 \leq i < j \leq m} \delta(C_i, C_j)}{max_{1 \leq k \leq m} \Delta_k}, \tag{5.1}$$

  where m is the number of clusters; C_i is a cluster of observations (vectors); $\delta(C_i, C_j)$ is an inter-cluster distance metric (an average distance between two clusters elements); and Δ_k is a cluster-size value describing average distance between observations in the cluster. Note that larger inter-cluster distances (better separation) and smaller cluster sizes (more compact clusters) lead to a higher value of DI;

- The *Fowlkes-Mallows index* [Fowlkes and Mallows, 1983] is an external evaluation measure (uses available grader labels) and is defined as follows:

$$FM = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}}, \tag{5.2}$$

  where TP stands for true positives, FP for false positives, and FN for false negatives. The index ranges from 0 to 1. A higher index indicates a higher similarity between a clustering and a benchmark classification.

Supervised learning accuracy

We use two widely used performance measures: exact agreement and quadratic weighted Kappa (for details see Section 4.1.2).

5.3.3 *Evaluation protocol and libraries*

In our experiments, we compare the accuracy obtained using our two-level clustering approach with the *baseline accuracy* of a prediction model that is built on the initial dataset. Figure 5.3 illustrates the workflow of our evaluation. It shows how we compare the baseline accuracy (left side of the figure) to the *joint prediction accuracy* obtained from several models built on the clustered dataset (right side of the figure).

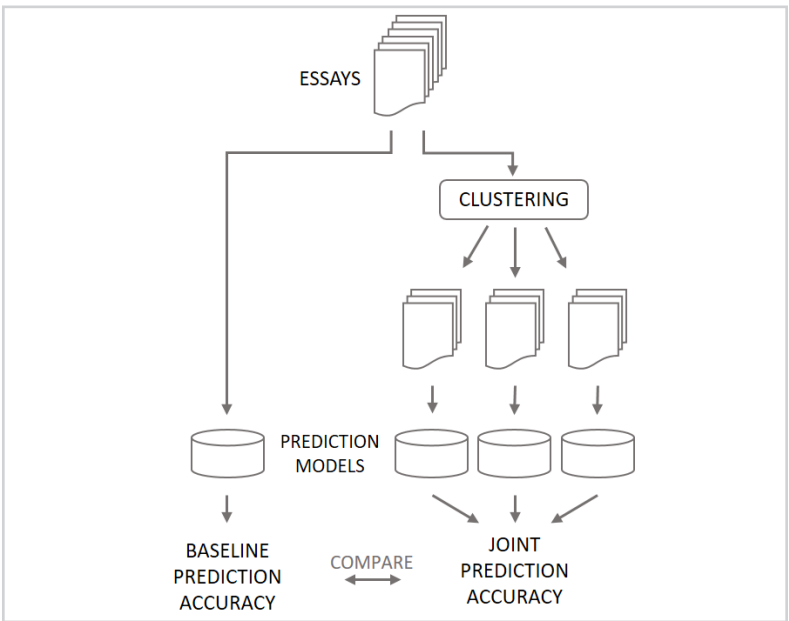

Figure 5.3

Comparing the baseline prediction accuracy (left) with the joint prediction accuracy of the clustered dataset (right).

The evaluation is performed using 10-fold cross-validation. Each training fold is used to perform all steps of the proposed approach: PCA, clustering, explanation computation, and model learning. Each test fold contains essays that are transformed into the existing PCA space defined by the training examples. After computing the

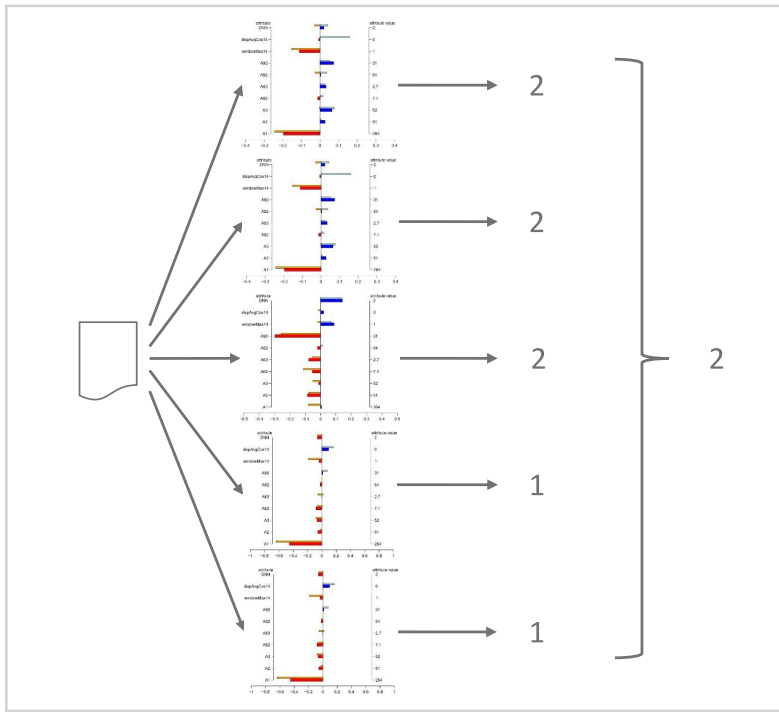

Figure 5.4

Predicting a score for an essay: We first obtain five nearest essays and their associated clusters on the second level. We illustrate the five models with their attributes' contributions. In the shown example two of the nearest essays were from the same cluster. The score represents the average prediction of all models.

explanations of predictions for test examples, they are mapped into the appropriate clusters. For each test example, we predict a score that is an average prediction of clusters' models of the five nearest essays (based on Euclidean distance). An example of calculating the prediction for an essay in dataset 3 is shown in Figure 5.4 where the models are represented with their attributes' contributions (in a form of nomograms). To ensure higher stability of results, which are dependent on the underlying clustering method, we repeat the clustering process using three different k-means clustering algorithms [Hartigan and Wong, 1979; Lloyd, 1957; Macqueen, 1967] and average their scores. To compare the significance of the difference between two prediction accuracies, we use the Wilcoxon signed-rank test [Kanji, 2006].

In our preliminary experiments on a subset of dataset 1, we used two methods for the transformation of the attribute space: PCA and Independent Component Analysis

(ICA) [Hyvarinen, 2013]. The results were comparable, thus we decided to proceed with only one method, i.e. PCA.

For computing the prediction explanation, the *ExplainPrediction* package[1] in R was used. The random forest algorithm, which has shown to achieve the best results in combination with the used attributes [Zupanc and Bosnić, 2017a], was used from the *randomForest* package[2] in R. The parameter settings for the random forest algorithm were: 100 trees, sampling with replacement, the number of attributes randomly sampled as candidates at each split is $\frac{|attributes|}{3}$.

5.4 Results

After evaluating the clustering quality of both approach variants (described in Section 5.2), we evaluate the best performing method on real datasets. We used the following parameter settings for both variants (as defined in Section 5.2):

- `n.clust.l1`: 20 first-level clusters for similar essays;
- `n.clust.l2`: 2 or 3 (we experimented with both values) second-level clusters for similar grading logics within each first-level cluster;
- `n.components` (for variant 1): 7 components, which corresponds to approximately 50% of variance in our domains and represents a sensible trade-off between keeping as few components as possible and as much information as possible at the same time;
- `n.import.attr` (for variant 1): 10 most important attributes;
- `n.infl.attr` (for variant 2): 10 most influential attributes.

Only for illustration purposes, we visualized the first-level clustering process on the 2D plot of PC1 (the most important principal component) vs. PC2 (the second most important principal component) on the dataset 2b using variant 1. Figure 5.5 shows the results of the k-means [Macqueen, 1967] clustering using 7 principal components. Although we cluster into 20 clusters (`n.clust.l1`) on the first level, the figure allows us to identify two larger non-overlapping groups of essays. In the following section we

[1] https://cran.r-project.org/web/packages/ExplainPrediction/index.html
[2] http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

proceed with a quantitative evaluation of clustering, focusing on the correspondence between the discovered second-level clusters and grader IDs.

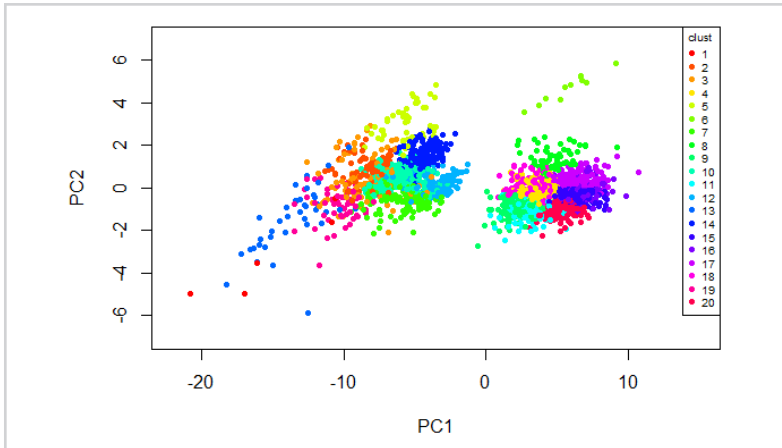

Figure 5.5

K-means clustering of essays into 20 clusters using 7 principal components.

5.4.1 *Experiments on artificial datasets*

In our artificial datasets we experimented with two and three clusters on the second level of clustering (parameter `n.clust.l2`) to simulate the frequent number of graders in our real datasets. Since our artificial datasets include two different graders (simulated by two versions of the LightSIDE system), the choice of clustering into two clusters also seems sensible. Given that the problem of distinguishing different graders is hard, we might also expect that the distinction between graders will be clear for some essay while for others not – hence the experiments with an additional third cluster, which could result in higher internal cluster homogeneity.

Clustering evaluation.

The results of internal (Dunn index) and external (Fowlkes-Mallows index) clustering evaluation are shown in Table 5.1. We can see that the indices do not report large differences between both variants. The Dunn index shows better results for Variant 2 and the Fowlkes-Mallows index shows slightly better results for Variant 1 when using three clusters on the second level. The values of the Fowlkes-Mallows index reach

Table 5.1

Clustering validation with the *Dunn index* and the *Fowlkes-Mallows index*, obtained on *artificial datasets* (ADSn). Indices are reported for all three k-means clustering methods: Hartigan and Wong - HW [Hartigan and Wong, 1979], Lloyd - Ll [Lloyd, 1957], and MacQueen - MQ [Macqueen, 1967].

appr.	measure	#C	method	ADS1	ADS2a	ADS2b	ADS3	ADS4	ADS5	ADS6	ADS7	ADS8	average
Variant 1	Dunn index	2	MQ	0.4286	0.6083	0.3493	0.5448	0.4170	0.4215	0.3789	0.2139	0.5449	*0.4341*
			Ll	0.4281	0.6299	0.3562	0.5273	0.4164	0.4006	0.3732	0.1657	0.5812	*0.4310*
			HW	0.4324	0.6339	0.3440	0.5522	0.4198	0.4182	0.5024	0.5481	0.5737	*0.4916*
			average	*0.4297*	*0.6240*	*0.3498*	*0.5415*	*0.4177*	*0.4134*	*0.4182*	*0.3092*	*0.5666*	0.4522
		3	MQ	0.3503	0.5444	0.3361	0.5096	0.2130	0.4463	0.3776	0.2141	0.4522	*0.3826*
			Ll	0.3619	0.5572	0.3431	0.5078	0.2197	0.4131	0.3568	0.1765	0.4541	*0.3767*
			HW	0.3548	0.5639	0.3347	0.5647	0.3454	0.4188	0.4583	0.1771	0.5168	*0.4149*
			average	*0.3557*	*0.5552*	*0.3380*	*0.5274*	*0.2593*	*0.4260*	*0.3975*	*0.1892*	*0.4744*	0.3914
	Fowlkes Mallows index	2	MQ	0.7076	0.7151	0.7008	0.7148	0.7218	0.7033	0.7034	0.7057	0.7214	*0.7104*
			Ll	0.7064	0.7179	0.7024	0.7158	0.7190	0.7051	0.7027	0.7055	0.7256	*0.7111*
			HW	0.7049	0.7144	0.7021	0.7135	0.7225	0.7047	0.7010	0.7118	0.7269	*0.7113*
			average	0.7063	0.7158	0.7018	0.7147	0.7211	0.7043	0.7024	0.7077	0.7246	*0.7110*
		3	MQ	0.7170	0.7277	0.7066	0.7323	0.7483	0.7051	0.7047	0.7155	0.7360	*0.7215*
			Ll	0.7178	0.7287	0.7063	0.7239	0.7448	0.7110	0.7043	0.7166	0.7331	*0.7207*
			HW	0.7235	0.7238	0.7101	0.7259	0.7482	0.7075	0.7071	0.7159	0.7392	*0.7223*
			average	*0.7194*	*0.7267*	*0.7076*	*0.7273*	*0.7471*	*0.7079*	*0.7053*	*0.7160*	*0.7361*	0.7215
Variant 2	Dunn index	2	MQ	0.4061	0.5529	0.4751	0.5703	0.5592	0.5681	0.3826	0.4782	0.2832	*0.4751*
			Ll	0.4276	0.5330	0.4765	0.5648	0.5522	0.5413	0.3828	0.4516	0.3181	*0.4720*
			HW	0.3922	0.5808	0.4766	0.5985	0.5857	0.5858	0.4027	0.4554	0.3083	*0.4873*
			average	*0.4086*	*0.5556*	*0.4761*	*0.5779*	*0.5657*	*0.5650*	*0.3893*	*0.4617*	*0.3032*	0.4781
		3	MQ	0.3115	0.4987	0.4097	0.5086	0.5287	0.4558	0.3218	0.3700	0.2599	*0.4072*
			Ll	0.2865	0.5062	0.4010	0.5120	0.4906	0.4367	0.3153	0.3650	0.2874	*0.4001*
			HW	0.3309	0.4679	0.4304	0.5558	0.5392	0.4819	0.3254	0.3743	0.2918	*0.4219*
			average	*0.3096*	*0.4909*	*0.4137*	*0.5255*	*0.5195*	*0.4581*	*0.3208*	*0.3697*	*0.2797*	0.4097
	Fowlkes Mallows index	2	MQ	0.7028	0.6994	0.6987	0.6971	0.7036	0.7036	0.6979	0.7078	0.7231	*0.7038*
			Ll	0.7028	0.6981	0.6985	0.6988	0.7015	0.7029	0.6973	0.7076	0.7226	*0.7033*
			HW	0.7025	0.6994	0.6943	0.6958	0.7047	0.7045	0.6977	0.7088	0.7225	*0.7034*
			average	*0.7027*	*0.6990*	*0.6972*	*0.6972*	*0.7033*	*0.7037*	*0.6976*	*0.7081*	*0.7227*	0.7035
		3	MQ	0.7089	0.7046	0.7098	0.7046	0.7128	0.7128	0.7075	0.7111	0.7311	*0.7115*
			Ll	0.7111	0.7082	0.7033	0.7066	0.7124	0.7081	0.7074	0.7125	0.7284	*0.7109*
			HW	0.7082	0.7077	0.7050	0.7067	0.7091	0.7073	0.7089	0.7114	0.7275	*0.7102*
			average	*0.7094*	*0.7068*	*0.7060*	*0.7060*	*0.7114*	*0.7094*	*0.7080*	*0.7117*	*0.7290*	0.7109

#C = number of clusters

over 0.7 (maximum is 1), meaning that we are able to determine that the clusters are well-formed. We proceed with the evaluation of the prediction accuracy.

Prediction accuracy evaluation

The results are shown in Table 5.2. They show that Variant 1 on the average performs significantly better than the baseline model when using three clusters on the second level. Furthermore, the results show a significant improvement in Kappa on five out of nine datasets (ADS2a, ADS2b, ADS4, ADS7 and ADS8) and does not significantly worsen on any dataset. When we compare Variant 2 with the baseline results using three clusters on the second level, we can observe that the Kappas significantly improve

on three datasets (ADS2a, ADS2b, and ADS4) and worsen on 1 dataset (ADS1); on the average, the approach does not perform significantly different than the baseline. However, all the results improve on average over all datasets.

Table 5.2

Quadratic weighted Kappa and *exact agreement* (EA) obtained on *artificial datasets* (ADSn) using the baseline, Variant 1, and Variant 2 approaches. Red color indicates higher and gray color lower accuracy in comparison to the baseline, while star (★) indicates significant difference ($p < 0.05$).

approach	measure	#C	ADS1	ADS2a	ADS2b	ADS3	ADS4
Baseline	Kappa	/	0.5546	0.4217	0.4091	0.4671	0.5687
	EA		0.4966	0.6192	0.6263	0.6216	0.6228
Variant 1	Kappa	2	0.5504	0.4338★	0.4262★	0.4698	0.5772
		3	0.5597	0.4504★	0.4318★	0.4737	0.5801★
	EA	2	0.4868	0.6292★	0.6408★	0.6386★	0.6343★
		3	0.4899	0.6371★	0.6471★	0.6369★	0.6364★
Variant 2	Kappa	2	0.5392★	0.4398★	0.4333★	0.4670	0.5762
		3	0.5419★	0.4468★	0.4390★	0.4684	0.5792★
	EA	2	0.4916	0.6317★	0.6496★	0.6240	0.6330
		3	0.4882	0.6321★	0.6521★	0.6221	0.6373★

#C = number of clusters

approach	measure	#C	ADS5	ADS6	ADS7	ADS8	average
Baseline	Kappa	/	0.7578	0.5945	0.5602	0.6835	0.5575
	EA		0.6804	0.6058	0.1363	0.1245	0.5037
Variant 1	Kappa	2	0.7566	0.6016	0.5623	0.6835	0.5613
		3	0.7549	0.6023	0.5746★	0.6997★	0.5697★
	EA	2	0.6812	0.6238★	0.1373	0.1286	0.5110★
		3	0.6808	0.6246★	0.1427	0.1297	0.5139★
Variant 2	Kappa	2	0.7576	0.5883	0.5660	0.6685★	0.5596
		3	0.7577	0.5856	0.5747	0.6761	0.5633
	EA	2	0.6879	0.6217★	0.1318	0.1235	0.5105
		3	0.6879	0.6212	0.1313	0.1266	0.5108

#C = number of clusters

To summarize the accuracy results, the Variant 1 produces better average results than Variant 2. Also, using three clusters yields better average results, regardless of the used variant. We can ascribe the reason to the fact that the distinction between graders is clear for some essays while not for the others (see Section 5.4.1). Based on this we choose Variant 1 for further evaluation on the real-world datasets.

Graders' profiles

The obtained clusters on the second level combine similar graders, meaning that each of `n.clust.l2` clusters on the second level represents a different grader profile. To evaluate how much those profiles differ, we plotted three example attributes' contributions (nomograms) that represent characteristics of each of `n.clust.l2` grader profiles in Figure 5.6. The presented models were built on clusters obtained on the ADS3 using k-means clustering [Lloyd, 1957] with 20 clusters on the first level and 3 clusters on the second level (we present an example for one of the 20 clusters only).

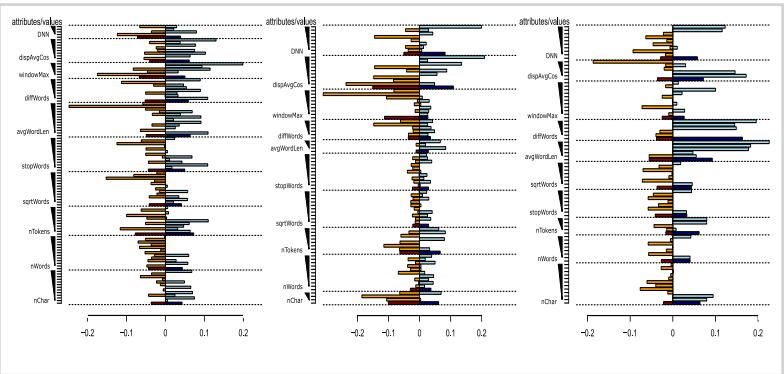

Figure 5.6

The attributes' contributions (in a form of nomograms) representing the grading characteristics for three different graders' profiles.

We can notice that the attribute value contributions between the three grader profiles differ. For example, different graders differently treat the length of an essay (number of characters – the attribute at the bottom): the first grader (left) has a tendency to treat long essays negatively and the average-sized essays positively, the second grader (middle) treats only long essays positively, and the third grader (right) treats only the short essays positively.

5.4.2 *Experiments on real-world datasets*

We proceed by further evaluating Variant 1 on the real datasetss. Since each essay from the real dataset can be graded either with scores from single graders or with the resolved score (aggregated from multiple graders), we evaluate Variant 1 in both scenarios, as well.

Single scores

Table 5.3 displays a comparison between the baseline and the joint prediction accuracies, when single scores are used. The results show that Variant 1 significantly improves the average quadratic weighted Kappa (but not also the exact agreement) when used with three clusters on the second level. When using three clusters on the second level, quadratic weighted Kappa (exact agreement) of Variant 1 significantly increases on three (four) of nine datasets (DS1, DS3, DS7; (DS1, DS3, DS4, DS5)), while it does not significantly decrease on any dataset.

Table 5.3

Quadratic weighted Kappa and *exact agreement* (EA) obtained on *real-world datasets with available single-score* (DSn) using the baseline and Variant 1 approaches. Red color indicates higher and gray color lower accuracy in comparison to the baseline, while star (★) indicates significant difference ($p < 0.05$).

approach	measure	#C	DS1	DS2a	DS2b	DS3	DS4
Baseline	Kappa	/	0.6091	0.6864	0.6521	0.6583	0.6622
	EA		0.3487	0.6996	0.6908	0.6717	0.6016
Variant 1	Kappa	2	0.6270★	0.6804	0.6444	0.6678	0.6631
		3	0.6416★	0.6851	0.6490	0.6707★	0.6680
	EA	2	0.3647★	0.6946	0.6842	0.6809	0.6203★
		3	0.3647★	0.6981	0.6917	0.6835★	0.6199★

#C = number of clusters

approach	measure	#C	DS5	DS6	DS7	DS8	average
Baseline	Kappa	/	0.7774	0.7249	0.4687	0.6665	0.6562
	EA		0.6467	0.6188	0.1299	0.1892	0.5108
Variant 1	Kappa	2	0.7805	0.7210	0.4727	0.6670	0.6582
		3	0.7859	0.7315	0.4882★	0.6697	0.6655★
	EA	2	0.6534	0.6104	0.1279	0.1841	0.5134
		3	0.6567★	0.6256	0.1289	0.1881	0.5175

#C = number of clusters

Resolved scores

Table 5.4 compares the baseline prediction accuracies with the joint prediction accuracies, when resolved scores are used. The results enable us to draw similar conclusions as on the datasets with single-grader scores, as follows. Variant 1 significantly improves the average quadratic weighted Kappa (but not also the exact agreement) when used

with three clusters on the second level. When using three clusters on the second level, quadratic weighted Kappa (exact agreement) of Variant 1 significantly increases on three (three) of nine datasets (DS3, DS6, DS8), while does not significantly worsen on any dataset.

Table 5.4

Quadratic weighted Kappa and *exact agreement* (EA) obtained on *real-world datasets with available resolved-score* (DSn) with available *resolved scores* using the baseline and Variant 1 approaches. Red color indicates higher and gray color lower accuracy in comparison to the baseline, while star (★) indicates significant difference ($p < 0.05$).

approach	measure	#C	DS1	DS2a	DS2b	DS3	DS4
Baseline	Kappa	/	0.8418	0.6864	0.6521	0.6789	0.6974
	EA		0.5507	0.6996	0.6908	0.6718	0.6165
Variant 1	Kappa	2	0.8472	0.6804	0.6443	0.6917★	0.6896
		3	0.8480	0.6812	0.6460	0.6920★	0.6897
	EA	2	0.5582	0.6946	0.6842	0.6888★	0.6105
		3	0.5582	0.6981	0.6917	0.6888★	0.6176

#C = number of clusters

approach	measure	#C	DS5	DS6	DS7	ADS8	average
Baseline	Kappa	/	0.8115	0.7307	0.7241	0.78185	0.7339
	EA		0.6920	0.6358	0.1478	0.1496	0.5394
Variant 1	Kappa	2	0.8142	0.7428★	0.7212	0.7776	0.7343
		3	0.8142	0.7442★	0.7292	0.8010★	0.7384★
	EA	2	0.6904	0.6450	0.1542	0.1558	0.5424
		3	0.6933	0.6479★	0.1502	0.1621★	0.5453

#C = number of clusters

We focused on a problem of learning from datasets that contain essays graded by multiple different graders. Due to the subjective nature of humans, this affects the learning algorithm as it has to model a noisy dependency between attributes and the grade. We proposed an approach for separating a set of essays into subsets that feature similar grading logics. We used a two-level clustering approach and employed the explanation methodology. The results show that we can significantly improve the average prediction accuracy by detecting groups of graders with similar grading characteristics. The essential step is the extraction of the crucial information using the explanation methodology, which is able to detect diverse grading logics. The higher joint prediction accuracy of the models describing distinctive graders is then plausible, comparing

to the model that has to combine different grading logics.

Moreover, the proposed methodology is robust and can work on top of any existing AEE system to increase the prediction accuracy of the assessments.

6

Conclusion

6.1 Main contributions to science

In this section we briefly summarize the main contributions to science listed in Section 1. With each contribution we list the sections where the topic is discussed. In addition, we also list our publications that discuss the topic. Note that the listed references were published in reviewed scientific journals or presented on international conferences and were thus internationally reviewed and discussed.

1. *New semantic attributes for evaluating semantic coherence of the text.* We present two groups of coherence attributes: spatial attributes described in Section 3.2.1 and network attributes described in Section 3.2.3. The proposed attributes allow us to evaluate one of the aspects of essay semantics, improve the prediction accuracy of the implemented AEE system and obtain state-of-the-art results. The proposed attributes are described in [Zupanc and Bosnić, 2014, 2017a; Zupanc et al., 2017].

2. *Methodology for cross-referencing facts in text with external fact sources.* We propose a system for automatic detection of semantic errors in an essay in Section 3.3.1. To implement the system we use entity recognition, coreference resolution, open information extraction, ontologies, and logic reasoner. The output of the system are three new semantic attributes and a semantic feedback. We propose the system SAGE and demonstrate its contributions in [Zupanc and Bosnić, 2017a].

3. *Methodology for detection of different graders.* Within the Section 5 we propose an approach for separating a set of essays into subsets that represent different graders. We use an explanation methodology and clustering to separate essay datasets. The results show that learning from the ensemble of separated models significantly improves the average prediction accuracy on artificial and real-world datasets. We describe the details in [Zupanc and Bosnić, 2017b].

6.2 Future research directions

The open challenges for our future work are scattered over different approaches we used through the development of the proposed AEE system. We mentioned a number of them already through the thesis and we summarize them in the next paragraph.

First and foremost, we shall further develop different semantic attributes and improve the semantic feedback. We shall test alternative approaches to TF-IDF for transforming text into attribute space to discover how the alternatives impact the results. Word or paragraph embeddings is an NLP technique where words or passages of text are mapped to vectors of real numbers. The literature describes approaches, such as Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014]. We shall represent parts of an essay with vectors using the above algorithms to determine if different representations could improve the results. Considering the network coherence attributes, we shall include a different approach for building networks by connecting sentences in the decreasing order of the computed similarities until the network becomes a connected graph (i.e. a graph with a single connected component). Furthermore, we shall upgrade our automatic error detection system and use other external sources to determine if statements in an essay are true and consistent. Moreover, we also want to develop and incorporate new approaches for unsupervised taxonomy learning, since our current approach uses only WordNet as the underlying taxonomy. One of the future goals is to incorporate inference rules to the ontology that will help detect implicit errors and facts/relations that are not explicitly written in an essay.

The performance of our system is dependent on a set of already graded essays. Other systems report on needing at least between 100 and 300 essays [Page, 1994; Landauer et al., 2000; Rudner et al., 2006; Rich et al., 2013] to build a scoring model. The smallest training set we used consisted of 650 essays and performed well. Future work shall include experiments about the influence of the training set size on the model's performance.

Development of the AEE field is of a great importance to teachers and students. It can not only help reduce teachers' load but can also helps students become more autonomous during their learning process. But currently, the majority of the systems only work for English essays. Thus, one of our goals for the future is also a development of the AEE system for Slovenian language. In recent years, a lot of NLP tools for Slovenian language were developed as part of national research projects. We still do not have all the tools we need (i.e. open information extractor), but we can start with a simple grading system that does not include the semantic feedback and further improve it when possible.

6.3 Final thoughts

The field of AEE has been developed to the point where the systems can accurately reproduce the human scores. Using deep learning, we might still achieve slightly higher reproduction accuracy even without considering semantics, yet due to the black box nature of deep learning, researchers cannot explain which aspects of the essay quality influence the final score. Nevertheless, the development is now striving towards the technologies that would help systems to understand the written text. Despite the recent advances in the field of artificial intelligence (AI), natural language understanding is still considered as the AI-hard problem [Yampolskiy, 2013], which is the biggest limitation for developing an AEE system that would work perfectly. However, existing NLP tools allow AEE systems to detect certain semantic errors in the written text.

To conclude, AEE systems with feedback can be an aid, not a replacement, for classroom instructions and can help students to achieve progress faster. Students can use SAGE in the classroom as well as at home, while learning. Feedback for each specific response returned by our system provides information on the quality of different aspects of writing, a score and a descriptive feedback. The system's constant availability for scoring gives a possibility to students to repetitively practice their writing at any time. SAGE is consistent as it predicts the same score for a single essay each time that essay is input to the system. This is important since the scoring consistency between prompts turned out to be one of the most difficult psychometric issues in human scoring [Attali, 2013]. Advantages of automated feedback are its anonymity, instantaneousness, and encouragement for repetitive improvements by giving students more practice in writing essays [Weigle, 2013]. By publicly providing the technical details and results of our AEE system, we also aim to promote the openness of this research field. Hopefully, this shall open opportunities to progress and help bring more AEE systems into practical applications.

Razširjeni povzetek

A

A.1 Uvod

Eseji veljajo za najboljše orodje za pisno preverjanje posameznikovega znanja. Študentom nudijo priložnost, da pokažejo široko paleto spretnosti in znanj, vključno z miselnimi sposobnostmi, kot sta sinteza in analiza [Valenti et al., 2003]. Kljub temu pa je ocenjevanje esejev ena izmed najbolj zamudnih, napornih in dragih dejavnosti za izobraževalne ustanove, ki učitelje obremenjuje z urami ocenjevanja pisnih izdelkov. Posledično dodeljujejo učencem manj pisnih nalog, s tem pa jim omejujejo prepotrebne izkušnje za doseganje učnih ciljev glede pisanja. Če želimo, da učenci in študenti postanejo boljši pisci, morajo pisati več, saj le s tem vadijo svoje spretnosti [Page, 1966].

Praktična rešitev za številne težave, povezane z ročnim ocenjevanjem, so avtomatski sistemi za ocenjevanje esejev. Po definiciji je avtomatsko ocenjevanje esejev (AOE) *postopek ocenjevanja in točkovanja pisne proze z računalniškim programom* [Shermis and Burstein, 2003]. Avtomatsko ocenjevanje esejev je multidisciplinarno področje, ki vključuje raziskave s področij računalništva, kognitivne psihologije, jezikoslovja in pisanja [Shermis and Burstein, 2013]. Ta avtomatski proces postaja zaželen način ocenjevanja v izobraževalnih ustanovah in tudi pri ocenjevanju standardiziranih testov.

Eden od glavnih problemov sistemov za avtomatsko ocenjevanje esejev je problem ocenjevanja semantične pravilnosti besedila. Obstoječi sistemi to večinoma rešujejo s primerjavo besedišča med novim esejem in z že ocenjenimi eseji ter z oceno uporabe diskurznih elementov. Nekateri sistemi pa uporabljajo metode, kot so latentna semantična analiza [Landauer et al., 1998], latentna Dirichletova alokacija [Kakkonen et al., 2008], in analiza vsebinskih vektorjev [Attali, 2011]. Poskusi kažejo, da lahko skladenjske in pomenske strukturne informacije bistveno izboljšajo učinkovitost modelov za avtomatsko ocenjevanje esejev. Zaenkrat pa le dva sistema [Gutierrez et al., 2014; Brent et al., 2010] uporabljata metode, ki vsaj delno preverjajo konsistentnost dejstev, zapisanih v eseju. Kljub trudu in izboljšavam pa omenjena sistema nista popolnoma avtomatska, saj zahtevata ročni vnos uporabnika pri oblikovanju povezav med pridobljenimi entitetami v eseju in entitetami v ontologiji.

Problem področja, o katerem v zadnjem času govori vse več znanstvenikov [Bejar, 2011; Attali, 2013; Williamson et al., 2012], je tudi golo reproduciranje ocen pristranskih ocenjevalcev. Raziskovalci večinoma privzamejo, da so ocene učiteljev objektivne in nepristranske, kar pa v realnosti ni tako. Učitelji in strokovnjaki ocenjujejo nekonsistentno in pristransko zaradi različnih karakteristik, kot so na primer izkušnje, interno

zaznavanje in okolje. Zaželeno je, da lahko sistemi za AOE zaznavajo določene napake v esejih neodvisno od človeških ocen in so tudi sposobni nuditi povratno informacijo o teh napakah.

Do nedavnega je bilo pomanjkanje javno dostopnih AOE sistemov, ki bi omogočili vpogled v metodologijo ocenjevanja, ena od glavnih ovir za dosego napredka na tem področju. Skoraj vse raziskave na področju avtomatskega ocenjevanja esejev so opravili v komercialnih ali profitnih družbah, ki so zaščitile svoje naložbe z omejevanjem dostopa do tehnoloških podrobnosti. Rudner je razvil prvi javno dostopen sistem, imenovan Bayesian Essay Test Scoring sYstem (BETSY) [Rudner and Liang, 2002]. Sistem BETSY je bil rezultat začetne raziskave glede uporabe Bayesovskega pristopa pri ocenjevanju esejev, avtorji pa nikoli niso nadaljevali s svojim delom. Kasneje sta Mayfield in Rosé predstavila LightSIDE [Mayfield and Penstein-Rosé, 2010], enostaven sistem za avtomatsko ocenjevanje z javno dostopno izvorno kodo. Ta program je zasnovan kot orodje za laike, ki lahko hitro izkoristijo tekstovno rudarjenje za različne namene, vključno z ocenjevanjem esejev. Pomanjkljivost sistema pa je, da vsebuje le preproste atribute, ki ne zaznavajo vseh karakteristik eseja. Poleg omenjenih sistemov je bilo v zadnjem času še kar nekaj poskusov, da bi področje avtomatskega ocenjevanja esejev naredili bolj odprto, med drugim tudi z objavo monografije avtorjev Shermis and Burstein [2013].

A.1.1 Prispevki k znanosti

Glavna tema disertacije je razvoj novega sistema za avtomatsko ocenjevanje esejev, ki naslavlja zgoraj opisane pomanjkljivosti obstoječih sistemov in področja.

1. *Novi semantični atributi za evalvacijo skladnosti ali koherence besedila.* Predlagamo dve skupini atributov skladnosti: prostorske atribute in atribute omrežij. Prvo skupino atributov razvijemo z opazovanjem semantičnih sprememb v toku besedila, drugo skupino atributov pa pridobimo z omrežji podobnosti povedi. Poleg tega razvijemo nov sistem za avtomatsko ocenjevanje esejev, ki uporablja nove atribute skladnosti. Njegovo delovanje podrobno analiziramo in ga primerjamo s primerljivimi sodobnimi sistemi. V povprečju sistem značilno izboljša napovedno točnost.

2. *Metodologija za primerjavo dejstev v besedilu z dejstvi iz zunanjih virov.* Predlagamo dodatne semantične atribute, ki na podlagi vsebine eseja zaznajo stavke,

ki so v nasprotju z resnico. *Avtomatski sistem za odkrivanje napak* (ASON), predstavlja novo metodologijo, ki odkriva semantične napake in nudi celovito povratno informacijo. Sistem samodejno preoblikuje besedilo eseja v ontologijo in ga primerja s predstavitvijo znanja iz zunanjih virov v obliki ontologije. Predlagani sistem za avtomatsko ocenjevanje esejev nadgradimo z novimi atributi in sistemom ASON, torej tudi z avtomatsko semantično povratno informacijo. Vse tehnološke podrobnosti sistema smo objavili v več člankih in jih predstavili na mednarodnih konferencah.

3. *Metodologija za odkrivanje različnih ocenjevalcev.* Predlagamo novo metodologijo za ločevanje množice podatkov, ki vsebuje ocene več različnih ocenjevalcev, v manjše podmnožice. Te manjše podmnožice vsebujejo le eseje, ki jih je ocenil isti ocenjevalec. Za razlikovanje med različnimi ocenjevalci uporabimo metodologijo razlage napovedi in gručenje razlag, ki nam omogoča detekcijo različnih odvisnosti (subjektivnih kriterijev ocenjevanja) med atributi eseja in njegovo oceno. Z eksperimenti pokažemo, da je model, ki se uči na ocenah večih ocenjevalcev, v povprečju slabši od ansambla modelov, ki predstavljajo različne ocenjevalce. Predlagan sistem za avtomatsko ocenjevanje esejev nadgradimo s predlagano metodologijo in v povprečju značilno izboljšamo napovedno točnost.

A.2 Sistem za avtomatsko ocenjevanje esejev SAGE

V disertaciji razvijemo sistem za avtomatsko ocenjevanje esejev z motivacijo, da naj sistem (1.) izboljša napovedno točnost in (2.) zagotovi avtomatsko semantično povratno informacijo. Nov sistem smo razvili v štirih fazah:

1. *Avtomatski ocenjevalec esejev – Automated Grader for Essays (AGE)*: sistem z jezikovnimi in primerjalnimi vsebinskimi atributi,

2. *Avtomatski ocenjevalec esejev+ –Automated Grader for Essays+ (AGE+)*: sistem AGE, nadgrajen z dodatnimi atributi skladnosti,

3. *Semantični avtomatski ocenjevalec esejev- – Semantic Automated Grader for Essays- (SAGE-)*: sistem AGE, nadgrajen z dodatnimi atributi konsistence in avtomatsko povratno informacijo.

4. *Semantični avtomatski ocenjevalec esejev – Semantic Automated Grader for Essays (SAGE)*: sistem AGE+, nadgrajen z dodatnimi atributi konsistence in avtomatsko povratno informacijo.

A.2.1 Avtomatski ocenjevalec esejev (AGE)

Razvili smo osnovni sistem za avtomatsko ocenjevanje esejev, ki temelji predvsem na atributih, opisanih v literaturi, in ga poimenovali AGE. Uporabili smo 72 skladenjskih atributov in jih razdelili v dve skupini:

- *Jezikovni atributi* opisujejo leksikalno sofisticiranost, slovnico in mehaniko eseja. Atribute pridobimo s pomočjo štetja besed, dolgih besed, različnih besed, uporabo stavčnih členov itd. Kompleksnejši atributi merijo stopnjo berljivosti, leksikalno različnost in napake pri črkovanju ali uporabi ločil in velike začetnice.

- *Vsebinski atributi* temeljijo na primerjavi neocenjenega eseja z že ocenjenimi. Za pridobitev teh atributov smo najprej združili podobno ocenjene eseje, nato pa smo primerjali vsebino novega eseja z vsebino že ocenjenih esejev.

A.2.2 Avtomatski ocenjevalec esejev+ (AGE+)

Predlagani sistem AGE smo nadgradili z dvema skupinama atributov skladnosti:

- *Atributi skladnosti, pridobljeni v visoko dimenzionalnem semantičnem prostoru.* Atribute smo gradili na domnevi, da se semantika skladnega eseja postopno spreminja skozi besedilo. Vsakega izmed esejev smo najprej razdelili v zaporedne prekrivajoče dele in jih z uporabo mere TF-IDF (frekvenca izraza v dokumentu - inverz frekvence izraza v zbirki dokumentov) preslikali v visokodimenzionalni semantični prostor. V tem prostoru smo merili različne karakteristike esejev s poudarkom na opazovanju semantičnih sprememb v toku besedila. Razvili smo 29 atributov, ki jih lahko razdelimo v tri skupine: osnovne mere skladnosti besedila (merijo razdaljo med deli eseja v semantičnem prostoru), prostorska skladnost (merijo centralno prostorsko tendenco in prostorsko razpršenost) in prostorska avtokorelacija (merijo stopnjo gručenja prostorskih podatkov) [Zupanc and Bosnić, 2014].

- *Atributi skladnosti, pridobljeni z uporabo omrežij podobnosti povedi.* Vsak esej smo spremenili v omrežje podobnosti povedi, kjer vsak stavek predstavlja eno

vozlišče v omrežju, povezave med vozlišči pa predstavljajo podobnost med dvema povedima. Razvili smo 32 atributov, ki jih lahko razvrstimo v tri skupine: osnovne strukturne metrike (osnovne karakteristike omrežja in indeksi povezanosti omrežja), sestavljene strukturne metrike (temeljijo na vsoti vrednosti povezav, pomnoženimi z relativno pomembnostjo povezave), in metrike omrežne entropije (nanašajo se na stopnjo vozlišč in njihovo distribucijo ter raznolikost) [Zupanc et al., 2017].

29 + 32 novih atributov skladnosti smo dodali sistemu AGE in zgradili nov sistem AGE+.

A.2.3 Semantični avtomatski ocenjevalec esejev (SAGE)

Nadaljevali smo z razvojem sistema za avtomatsko detekcijo semantičnih napak v eseju in predlagali popolnoma avtomatski sistem, ki izboljša AGE+: sistem odkriva semantične napake in nudi povratno informacijo. Sistem smo poimenovali SAGE - Semantic Automated Grader for Essays [Zupanc and Bosnić, 2017a]. Glavna novost je *avtomatski sistem za odkrivanje napak* (ASON).

ASON najprej zgradi temeljno ontologijo, ki je sestavljena iz:

- splošnega znanja – ontologije COSMO [Cassidy, 2009],
- domenskega znanja – domenska ontologija, ki zajema podrobno znanje domene, in
- izvornega besedila – znanja, ki ga ASON pridobi iz besedila, na podlagi katerega je esej napisan.

Vzporedno iz eseja zgradimo semantični graf, kar poteka v več korakih:

1. predprocesiranje,
2. odkrivanje koreferenčnosti med omenitvami in
3. ekstrakcija relacij med entitetami.

Rezultat ekstrakcije informacij so trojke `{<arg1,rel,arg2>}`, ki opisujejo relacije `rel` med argumenti (osebki ali predmeti) `arg1` in `arg2`. Te relacije nato iterativno dodajamo v temeljno ontologijo in vsakič s pomočjo avtomatskega logičnega misleca

(Hermit [Motik et al., 2009]) preverimo, če je ontologija konsistentna. Če zaznamo nekonsistentnost, nam sistem omogoča, da zaznamo tudi nekonsistenta stavka. Le-to nam omogoča, da lahko učencu vrnemo povratno informacijo, kje v eseju je prišlo do napake in kakšna je ta napaka.

Poleg avtomatske semantične povratne informacije nam ASON nudi tudi tri nove atribute, ki jih vključuimo v sistem SAGE – ti opisujejo število semantičnih napak v eseju.

A.3 Primerjava sistemov AGE, AGE+, SAGE- in SAGE s sodobnimi sistemi za avtomatsko ocenjevanje esejev

Trije predlagani sistemi predstavljajo tri različne vidike ocenjevanja:

- ocenjevanje eseja brez razumevanja vsebine (AGE),
- ocenjevanje vsebine skozi skladnost (AGE+),
- ocenjevanje konsistentnosti dejstev v eseju z zagotovljeno povratno informacijo (SAGE-) in
- ocenjevanje semantike eseja z zagotovljeno povratno informacijo (SAGE).

Za izgradnjo modelov na pridobljenih atributih smo uporabili metodo za izbiro atributov in model naključnih gozdov. Da bi ocenili kakovost predlaganih sistemov, smo jih primerjali tako med seboj kot tudi s sodobnimi sistemi za avtomatsko ocenjevanje esejev. Uporabili smo podatkovne množice esejev, dostopne na spletni strani Kaggle[1].

Za analizo potencialnih koristi predlaganih atributov smo najprej ocenili njihovo pomembnost. Ugotovili smo, da na končno oceno najbolj vplivata skupno število besed in število različnih besed ter tudi ocena, ki jo dobimo pri primerjavi vsebine z že ocenjenimi eseji. Trije izmed predlaganih atributov skladnosti pa so se v povprečju uvrstili med prvih deset najvplivnejših atributov. Poleg tega rezultati kažejo, da 38% v povprečju najvplivnejših atributov predstavljajo atributi skladnosti. Najbolje uvrščen predlagani atribut konsistentnosti se nahaja v povprečju na 46. mestu.

Pri primerjavi napovedne točnosti modelov AGE, AGE+, SAGE- in SAGE smo ugotovili, da se napovedna točnost v povprečju značilno zviša, ko sistem AGE nadgradimo

[1] Podatke se lahko pridobi na spletni strani Kaggle http://www.kaggle.com/c/asap-aes/data ali spletni strani ASAP http://www.scoreright.org/

z atributi skladnosti. Ko sistemu AGE+ dodamo še atribute konsistence, napovedna točnost ni značilno slabša na nobeni podatkovni množici, se pa napovedna točnost značilno poveča na 50% podatkovnih množicah. SAGE je poleg tega značilno boljši v povprečju na vseh datasetih in kot nadgradnjo vsebuje semantično povratno informacijo. Tudi sistem SAGE- v enem izmed eksperimentov pokaže signifikantno izboljšanje napovedne točnosti v primerjavi s sistemom AGE.

Sistem SAGE smo primerjali s sodobnimi sistemi za avomatsko ocenjevanje esejev. SAGE v povprečju dosega značilno višjo točnost napovedi kot skoraj vsi ostali sodobni sistemi in pristopi.

A.4 Avtomatsko ločevanje različnih ocenjevalcev

V zadnjem delu disertacije se lotevamo problema pristranskosti ocenjevalcev. Domnevamo, da je človeško ocenjevanje nekonsistentno in pristransko in da bodo modeli, zgrajeni na množicah esejev, pridobljenih z ločevanjem različnih ocenjevalcev, prinesli višjo napovedno točnost. Tako predlagamo nov pristop za avtomatsko ločevanje ocenjevalcev glede na njihove karakteristike ocenjevanja esejev. Za razlikovanje med ocenjevalci uporabljamo *metodo razlage napovedi* [Štrumbelj et al., 2009], ki nam omogoča, da zaznamo različne odvisnosti (subjektivne kriterije ocenjevanja) med atributi esejev in njegovo oceno.

Pri ločevanju ocenjevalcev se opiramo na hipotezo, da ocenjevalci z različnimi subjektivnimi kriteriji ocenjevanja podobne eseje ocenijo z drugačno oceno ali razlago ocene. Predlagani pristop lahko opišemo v petih korakih:

1. V prvem koraku zmanjšamo dimenzionalnost vhodnega podatkovnega prostora, ki predstavlja karakteristike eseja z uporabo analize glavnih komponent (PCA) [Abdi and Williams, 2010].

2. Nato izvedemo prvo gručenje, pri katerem združujemo podobne eseje na podlagi naše zgoraj opisane predpostavke. Gručenje izvedemo na lastnostih glavnih komponent eseja, ki smo jih pridobili iz atributov eseja.

3. Izračunamo *razlago napovedi* [Štrumbelj et al., 2009], ki odraža subjektivne kriterije ocenjevalca za vsak ocenjen esej.

4. Na drugem nivoju gručimo podobne ocenjevalce, pri tem pa si pomagamo z razlago napovedi, ki omogoča, da za vsakega ocenjevalca določimo, kako pomemb-

na je zanj posamezna karakteristika eseja. Tako pridobimo želene podmnožice, ki predstavljajo različne ocenjevalce.

5. V zadnjem koraku zgradimo napovedne modele za vsako pridobljeno podmnožico. V fazi ocenjevanja nato vsak nov, neocenjen esej razvrstimo v podmnožico, ki najbolj ustreza njegovim karakteristikam, in uporabimo pripadajoči napovedni model.

Za evalvacijo predlaganega pristopa smo uporabili naravne in umetne podatkovne množice. Za naravne podatkovne množice smo uporabili eseje, uporabljene tudi v poglavju A.3. Umetne podatkovne množice pa smo pridobili tako, da smo eseje is naravnih podatkovnih množic ocenili z dvema ocenjevalcema: dvema različicama sistemoma LightSIDE [Mayfield and Rosé, 2013] in naključno dodelili eno izmed ocen vsakemu eseju tako, da sta na koncu oba ocenjevalca ocenila enako število esejev.

Rezultati kažejo, da lahko s predlaganim pristopom ločimo ocenjevalce v skupine, ki predstavljajo različno subjektivne kriterije ocenjevanja in v povprečju značilno izboljšamo napovedno točnost. Bistven korak predlaganega pristopa je pridobivanje informacij z uporabo metodologije razlag napovedi, s pomočjo katere lahko zaznamo različne subjektivne kriterije ocenjevanja. Višja napovedna točnost ansambla modelov je zato razumljiva, saj jo primerjamo z napovedno točnostjo modela, ki mora modelirati različne subjektivne kriterije ocenjevanja. Poleg tega lahko predlagana metodologija deluje kot nadgradnja za vse sodobne sisteme za avtomatsko ocenjevanje esejev in pri tem izboljša napovedno točnost.

A.5 Zaključek

Zaključimo z mislijo, da so lahko sistemi za avtomatsko ocenjevanje esejev s povratno informacijo v veliko pomoč in ne nadomestilo za napotke in komentarje učiteljev. Predlagani sistem vrača semantično povratno informacijo in daje učencem možnost, da vadijo in izboljšujejo svoje pisanje,s kadar želijo. Z javno objavo vseh podrobnosti delovanja in rezultatov predlaganega sistema, tudi upamo, da bomo spodbudili razvoj tega področja.

BIBLIOGRAPHY

Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. doi: 10.1002/wics.101.

Helen B. Ajay, P. I. Tillet, and Ellis Batten Page. Analysis of essays by computer (AEC-II). Technical report, U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development, Washington, D.C., 1973. URL https://catalogue.nla.gov.au/Record/5258250.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic Text Scoring Using Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 715–725, Berlin, Germany, 2016. doi: 10.18653/v1/P16-1068.

Lucas Antiqueira, Maria G. V. Nunes, Osvaldo N. Oliveira, and Luciano da F. Costa. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, 373:811–820, 2007. doi: 10.1016/j.physa.2006.06.002.

Ioannis E. Antoniou and Eleni T. Tsompa. Statistical analysis of weighted networks. *Discrete Dynamics in Nature and Society*, 2008(Article ID 375452):16p, 2008. doi: 10.1155/2008/375452.

Yigal Attali. A Differential Word Use Measure for Content Analysis in Automated Essay Scoring. *ETS Research Report Series*, 36, 2011.

Yigal Attali. Validity and Reliability of Automated Essay Scoring. In Mark D. Shermis and Jill C. Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions2*, chapter 11, pages 181–198. Routledge, New York, 2013.

Yigal Attali and Jill Burstein. Automated Essay Scoring With e-rater V.2. *The Journal of Technology, Learning and Assessment*, 4(3):3–29, 2006.

Regina Barzilay and Mirella Lapata. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34, 2008. doi: 10.1162/coli.2008.34.1.1.

Hannah Bast and Elmar Haussmann. Open Information Extraction via Contextual Sentence Decomposition. In *Proceedings of the 7th International Conference on Semantic Computing (ICSC)*, pages 154–159, Irvine, California, 2013. doi: 10.1109/ICSC.2013.36.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language Reference. Technical report, 2004. URL https://www.w3.org/TR/owl-ref/.

Isaac I. Bejar. A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, 18(3):319–341, aug 2011. doi: 10.1080/0969594X.2011.555329.

Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, pages 294–303, Waikiki, Honolulu, Hawaii, 2008. doi: 10.3115/1613715.1613756.

Li Bin and Yao Jian-Min. Automated Essay Scoring Using Multi-classifier Fusion. *Communications in Computer and Information Science*, 233:151–157, 2011. doi: 10.1007/978-3-642-24010-2_21.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

Stefano Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006. doi: 10.1016/j.physrep.2005.10.009.

Edward Brent and Martha Townsend. Automated essay grading in the sociology classroom. In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences?*, chapter 13, pages 177–198. Utah State University Press, 2006.

Edward Brent, Curtis Atkisson, and Nathaniel Green. Time-shifted Collaboration: Creating Teachable Moments through Automated Grading. In A. A. Juan,

T. Daradournis, and S. Caballe, editors, *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Support*, pages 55–73. IGI Global, 2010. doi: 10.4018/978-1-60566-786-7.ch004.

Brent Bridgeman. Human Ratings and Automated Essay Evaluation. In Mark D. Shermis and Jill C. Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 13, pages 221–232. Routledge, New York, 2013.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. Computer Analysis of Essays. In *Proceedings of the NCME Symposium on Automated Scoring*, number April, pages 1–13, Montreal, 1998.

Jill Burstein, Martin Chodorow, and Claudia Leacock. Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine*, 25(3):27–36, 2004.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 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 ACL*, number June, pages 681–684, Los Angeles, California, 2010. Association for Computational Linguistics.

Jill Burstein, Joel Tetreault, and Nitin Madnani. The E-rater® Automated Essay Scoring System. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 4, pages 55–67. Routledge, New York, 2013a.

Jill C. Burstein, Joel R. Tetreault, Martin Chodorow, Daniel Blanchard, and Slava Andreyev. Automated Evaluation of Discourse Coherence Quality in Essay Writing. In Mark D. Shermis and Jill C. Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 16, pages 267–280. Routledge, New York, 2013b.

Patrick Cassidy. Toward an open-source foundation ontology representing the Longman's defining vocabulary: The COSMO ontology OWL version. In *Proceedings of the 3rd International Ontology for the Intelligence Community Conference*, pages 45–48, Fairfax, VA, 2009.

Yllias Chali and Sadid A. Hasan. On the Effectiveness of Using Syntactic and Shallow Semantic Tree Kernels for Automatic Assessment of Essays. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 767–773, Nagoya, Japan, 2013.

Danqi Chen and Christopher D. Manning. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 740–750, Doha, Qatar, 2014. doi: 10.3115/v1/D14-1082.

Hongbo Chen, Ben He, Tiejian Luo, and Baobin Li. A Ranked-Based Learning Approach to Automated Essay Scoring. In *Proceedings of the Second International Conference on Cloud and Green Computing*, pages 448–455. Ieee, nov 2012. doi: 10.1109/CGC.2012.41.

James R. Christie. Automated Essay Marking – for both Style and Content. In *Proceedings of the Third Annual Computer Assisted Assessment Conference*, 1999.

Philip J. Clark and Francis C. Evans. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology*, 35(4):445–453, 1954. doi: 10.2307/1931034.

Peter J. Congdon and Joy McQueen. The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2):163–178, 2000. doi: 10.1111/j.1745-3984.2000.tb01081.x.

Luciano Del Corro and Rainer Gemulla. ClausIE: Clause-Based Open Information Extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366, Rio de Janeiro, Brazil, 2013. doi: 10.1145/2488388.2488420.

Luciano da F. Costa, Francisco A. Rodrigues, Gonzalo Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007. doi: 10.1080/00018730601170527.

Michael A. A. Cox and Trevor F. Cox. Multidimensional Scaling. In *Handbook of Data Visualization*, pages 315–347. Springer, Berlin, Heidelberg, 2008.

Kinkar C. Das and Nenad Trinajstić. Relationship between the eccentric connectivity index and Zagreb indices. *Computers and Mathematics with Applications*, 62:1758–1764, 2011. doi: 10.1016/j.dam.2013.05.034.

William H. Dubay. *Smart Language: Readers, Readability, and the Grading of Text*. BookSurge Publishing, 2007.

J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974. doi: 10.1080/01969727408546059.

Thomas Eckes. *Rater types in writing performance assessments: A classification approach to rater variability*, volume 25. 2008. doi: 10.1177/0265532207086780.

Catherine Elder, Gary Barkhuizen, Ute Knoch, and Janet von Randow. Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1):37–64, 2007. doi: 10.1177/0265532207071511.

George Jr. Engelhard. Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2):93–112, 1994. doi: 10.1111/j.1745-3984.1994.tb00436.x.

Oren Etzioni, Michael Cafarella, and Michele Banko. Open Information Extraction, Patent, 2014. URL https://www.google.com/patents/US20140032209.

Anhar Fazal, Tharam Dillon, and Elizabeth Chang. Noise Reduction in Essay Datasets for Automated Essay Grading. *Lecture Notes in Computer Science*, 7046:484–493, 2011. doi: 10.1007/978-3-642-25126-9_60.

Peter W. Foltz. Discourse Coherence and LSA. In Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors, *Handbook of Latent Semantic Analysis*, chapter 9, pages 167–184. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 2007.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307, 1998. doi: 10.1080/01638539809545029.

Peter W. Foltz, Lynn A. Streeter, Karen E. Lochbaum, and Thomas K. Landauer. Implementation and Applications of the Intelligent Essay Assessor. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 5, pages 68–88. Routledge, New York, 2013.

Edward B. Fowlkes and Colin L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983. doi: 10.1080/01621459.1983.10478008.

Pablo Gamallo. An Overview of Open Information Extraction. In Maria João Varanda Pereira, José Paulo Lea, and Alberto Simões, editors, *Invited Talk at the 3rd Symposium on Languages, Applications and Technologies, SLATE'14*, pages 13–16, 2014.

Pablo Gamallo, Marcos Garcia, and Santiago Fernandez-Lanza. Dependency-Based Open Information Extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, Avignon, France, 2012.

Roy C. Geary. The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5(3):115–145, 1954. doi: 10.2307/2986645.

Arthur Getis and J. Keith Ord. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3):189–206, 1992. doi: 10.1111/j.1538-4632.1992.tb00261.x.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, mar 2006. doi: 10.1007/s10994-006-6226-1.

Corrado Gini. *Variabilità e Mutuabilità*. Tipografia di Paolo Cuppini, Bologna, Italy, 1912.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics*, 21(2): 203–226, 1995. doi: 10.21236/ADA324949.

Tom Gruber. Ontology. In Ling Liu and M. Tamer Ozsu, editors, *Encyclopedia of Database Systems*. Springer-Verlag, 2009. URL http://tomgruber.org/writing/ontology-definition-2007.htm.

Fernando Gutierrez, Daya C. Wimalasuriya, and Dejing Dou. Using information extractors with the neural electromagnetic ontologies. In *Proceedings of the 2011th Confederated international conference on On the move to meaningful internet systems (OTM11)*, pages 31–32, 2011. doi: 10.1007/978-3-642-25126-9_8.

Fernando Gutierrez, Dejing Dou, Stephen Fickas, and Gina Griffiths. Providing grades and feedback for student summaries by ontology-based information extraction. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*, pages 1722–1726, 2012. doi: 10.1145/2396761.2398505.

Fernando Gutierrez, Dejing Dou, Adam Martini, Stephen Fickas, and Hui Zong. Hybrid Ontology-based Information Extraction for Automated Text Grading. In *Proceedings of 12th International Conference on Machine Learning and Applications*, pages 359–364, 2013. doi: 10.1109/ICMLA.2013.73.

Fernando Gutierrez, Dejing Dou, Stephen Fickas, and Gina Griffiths. Online Reasoning for Ontology-Based Error Detection in Text. *On the Move to Meaningful Internet Systems: OTM 2014 Conferences Lecture Notes in Computer Science*, 8841:562–579, 2014. doi: 10.1007/978-3-662-45563-0_34.

Fernando Gutierrez, Dejing Dou, Nisansa de Silva, and Stephen Fickas. Online Reasoning for Semantic Error Detection in Text. *Journal on Data Semantics*, 6(3): 139–153, 2017. doi: 10.1007/s13740-017-0079-6.

John A. Hartigan and Manchek A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108, 1979. doi: 10.1890/11-0206.1.

Marti A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating Multiple Aspects of Coherence in Student Essays. In *Proceedings of HLT-NAACL*, pages 185–192, Boston, MA, 2004. ACL.

Derrick Higgins, Jill Burstein, and Yigal Attali. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(02):145–159, may 2006. doi: 10.1017/S1351324906004189.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

A. Hyvarinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534–20110534, 2013. doi: 10.1098/rsta.2011.0534.

Md. Monjurul Islam and A. S. M. Latiful Hoque. Automated essay scoring using Generalized Latent Semantic Analysis. *Journal of Computers*, 7(3):616–626, 2012. doi: 10.4304/jcp.7.3.616-626.

Tuomo Kakkonen, Niko Myller, Erkki Sutinen, and Jari Timonen. Comparison of Dimension Reduction Methods for Automated Essay Grading. *Educational Technology & Society*, 11(3):275–288, 2008.

Gopal K. Kanji. *100 Statistical Tests*. SAGE Publications, London, Thousand Oaks, New Delhi, 3rd edition, 2006.

Xiaohua Ke, Yongqiang Zeng, and Haijiao Luo. Autoscoring Essays Based on Complex Networks. *Journal of Educational Measurement*, 53(4):478–497, 2016. doi: 10.1111/jedm.12127.

Igor Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *Proceedings of European Conference on Machine Learning (ECML'94)*, pages 171–182, Catania, Italy, 1994. doi: 10.1007/3-540-57868-4_57.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, jan 1998. doi: 10.1080/01638539809545028.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. The Intelligent Essay Assessor. *IEEE Intelligent systems*, 15 (5):27–31, 2000.

George Leckie and Jo Anne Baird. Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4):399–418, 2011. doi: 10.1111/j.1745-3984.2011.00152.x.

Stuart P. Lloyd. Least squares quantization in PCM. Technical Report RR-5497. Technical report, Bell Lab, 1957.

Susan M. Lottridge, E. Matthew Schulz, and Howard C. Mitzel. Using Automated Scoring to Monitor Reader Performance and Detect Reader Drift in Essay Scoring. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 14, pages 233–250. Routledge, New York, 2013.

James Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967. doi: citeulike-article-id:6083430.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*, pages 55–60, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Oliver Mason and Ian Grove-Stephenson. Automated free text marking with paperless school. In *Proceedings of the Sixth International Computer Assisted Assessment Conference*, pages 213–219, 2002.

Elijah Mayfield and Carolyn Penstein-Rosé. An Interactive Tool for Supporting Error Analysis for Text Mining. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 25–28, Los Angeles, CA, 2010.

Elijah Mayfield and Carolyn Rosé. LightSIDE: Open Source Machine Learning for Text. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 8, pages 124–135. Routledge, New York, 2013.

Andrew Mellor. Essay Length, Lexical Diversity and Automatic Essay Scoring. *Memoirs of the Osaka Institute of Technology*, 55(2):1–14, 2011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems NIPS'13*, pages 3111–3119, Lake Tahoe, Nevada, USA, 2013. Curran Associates Inc. doi: 10.1162/jmlr.2003.3.4-5.951.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995. doi: 0.1145/219717.219748.

Eleni Miltsakaki and Karen Kukich. Automated Evaluation of Coherence in Student Essays. *Proceedings of LREC-2000, Linguistic Resources in Education Conf.,Athens, Greece, 2000*, pages 140–147, 2000.

Thomas M. Mitchell. *Machine learning*. McGraw-Hill, Inc., New York, NY, USA, 1997.

Patrick Alfred Pierce Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23, 1950.

Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009. doi: 10.1613/jair.2811.

Carol M. Myford and Edward W. Wolfe. Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4):371–389, 2009. doi: 10.1111/j.1745-3984.2009.00088.x.

Ellis B. Page. The Imminence of... Grading Essays by Computer. *Phi Delta Kappan*, 47(5):238–243, 1966.

Ellis Batten Page. Computer Grading of Student Prose, Using Modern Concepts and Software. *Journal of Experimental Education*, 62(2):127–142, 1994. doi: 10.1080/00220973.1994.9943835.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. A Joint Framework for Coreference Resolution and Mention Head Detection. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL'15)*, pages 12–21, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1002.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1532–1543, Doha, Qatar, 2014. ACL. doi: 10.3115/v1/D14-1162.

Les Perelman. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111, 2014. doi: 10.1016/j.asw.2014.05.001.

Vasin Punyakanok and Dan Roth. The Use of Classifiers in Sequential Inference. In *Proceedings of Neural Information Processing Systems (NIPS'01)*, pages 995–1001, Vancouver, British Columbia, 2001.

Ali Reza Rezaei and Michael Lovorn. Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1):18–39, 2010. doi: 10.1016/j.asw.2010.01.003.

Changhua S. Rich, M. Christina Schneider, and Juan M. D'Brot. Applications of Automated Essay Evaluation in West Virginia. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 7, pages 99–123. Routledge, New York, 2013.

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004. doi: 10.1108/00220410410560582.

Lawrence M. Rudner and Tahung Liang. Automated Essay Scoring Using Bayes' Theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3–21, 2002.

Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. An Evaluation of the IntelliMetric Essay Scoring System. *The Journal of Technology, Learning and Assessment*, 4(4):3–20, 2006.

Matthew T. Schultz. The IntelliMetric Automated Essay Scoring Engine - A Review and an Application to Chinese Essay Scoring. In Mark D. Shermis and Jill C. Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 6, pages 89–98. Routledge, New York, 2013.

Mark D. Shermis and Jill Burstein. Introduction. In Mark D. Shermis and Jill Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*, pages xiii–xvi. Lawrence Erlbaum Associates, Manwah, NJ, 2003.

Mark D. Shermis and Jill C. Burstein, editors. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, New York, 2013.

Mark D. Shermis and Ben Hamner. Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 19, pages 313–346. Routledge, New York, 2013.

Mark D. Shermis, Howard R. Mzumara, Jennifer Olson, and Susanmarie Harrington. On-line Grading of Student Essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, 26(3):247–259, 2001. doi: 10.1080/02602930120052404.

Mark D. Shermis, Jill Burstein, and Sharon Apel Bursky. Introduction to Automated Essay Evaluation. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 1, pages 1–15. Routledge, New York, 2013.

Christian Smith and Arne Jönsson. Automatic Summarization As Means Of Simplifying Texts , An Evaluation For Swedish. In Bolette Sandford Pedersen, Gunta Nešpore, and Inguna Skadina, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA'11)*, pages 198–205, 2011.

Malbert III Smith. The Reading-Writing Connection. Technical report, MetaMetrics, 2009. URL https://metametricsinc.com/research-publications/reading-writing-connection/.

Malbert III Smith, Anne Schiano, and Elizabeth Lattanzio. Beyond the classroom. *Knowledge Quest*, 42(3):20–29, 2014.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3): 647–665, 2014. doi: 10.1007/s10115-013-0679-x.

Erik Štrumbelj, Igor Kononenko, and Marko Robnik Šikonja. Explaining instance classifications with interactions of subsets of feature values. *Data and Knowledge Engineering*, 68(10):886–904, 2009. doi: 10.1016/j.datak.2009.01.004.

Kaveh Taghipour and Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pages 1882–1891, Austin, Texas, USA, 2016. doi: 10.18653/v1/D16-1193.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education*, 2:319–330, 2003. doi: 10.28945/331.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684): 440–442, 1998. doi: 10.1038/30918.

Sara C. Weigle. English as a Second Language Writing and Automated Essay Evaluation. In Mark D. Shermis and Jill C. Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter 3, pages 36–54. Routledge, New York, 2013.

Robert Williams and Heinz Dreher. Automatically Grading Essays with Markit©. *Issues in Informing Science and Information Technology*, 1:693–700, 2004. doi: 10.28945/769.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Isues and Practice*, 31(1): 2–13, 2012. doi: 10.1111/j.1745-3992.2011.00223.x.

Daya C. Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3): 306–323, 2010. doi: 10.1177/0165551509360123.

Fei Wu and Daniel S. Weld. Open Information Extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, 2010.

Hsiao-ping Wu and Esther V. Garza. Types and Attributes of English Writing Errors in the EFL Context—A Study of Error Analysis. *Journal of Language Teaching and Research*, 5(6):1256–1262, 2014. doi: 10.4304/jltr.5.6.1256-1262.

Roman V. Yampolskiy. Turing test as a defining feature of AI-completeness. In Xin-She Yang, editor, *Artificial Intelligence, Evolutionary Computation and Metaheuristics (AIECM) –In the footsteps of Alan Turing*, chapter 1, pages 3–17. Springer, London, 2013. doi: 10.1007/978-3-642-29694-9-1.

Yiting Yang and Linyuan Lu. The Randić index and the diameter of graphs. *Discrete Mathematics*, 311(14): 1333–1343, 2011. doi: 10.1016/j.disc.2011.03.020.

Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, and Neil Heffernan. A Memory-Augmented Neural Model for Automated Grading. In *Proceedings of the 4th ACM Conference on Learning @ Scale - L@S'17*, pages 189–192, Cambridge, Massachusetts, USA, 2017. ACM. doi: 10.1145/3051457.3053982.

Kaja Zupanc and Zoran Bosnić. Automated Essay Evaluation Augmented with Semantic Coherence Measures. In *2014 IEEE International Conference on Data Mining*, pages 1133–1138, Shenzhen, China, 2014. doi: 10.1109/ICDM.2014.21.

Kaja Zupanc and Zoran Bosnić. Advances in the field of automated essay evaluation. *Informatica*, 39(4):383–395, 2015.

Kaja Zupanc and Zoran Bosnić. Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, 120: 118–132, 2017a. doi: 10.1016/j.knosys.2017.01.006.

Kaja Zupanc and Zoran Bosnić. Improvement of automated essay grading by separation of different graders. *Under review*, 2017b.

Kaja Zupanc, Miloš Savić, Zoran Bosnić, and Mirjana Ivanović. Evaluating coherence of essays using sentence-similarity networks. In *Proceedings of the 18th International Conference on Computer Systems and Technologies: CompSysTech'17, ACM International Conference Proceeding Series 1369*, pages 65–72, Ruse, Bulgaria, 2017. doi: 10.1145/3134302.3134322.