Computer-based Assessment of Essays Based on Multiple Documents:

Evaluating the Use of Sources

Joseph P. Magliano[1], Peter Hastings[2], Kristopher Kopp[3], Dylan Blaum[2], & Simon Hughes[2]

1. Northern Illinois University
2. DePaul University
3. Arizona State University

To appear in J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.) *Handbook of multiple source use*. New York, NY: Routledge.

Many personal, professional, and academic literacy activities require one to read multiple documents (texts, graphs, videos), extract content, and represent it in an integrated mental model, which has been referred to as a *documents model* (OECD, 2008; Rouet & Britt, 2011; Strømsø, Bråten, & Britt, 2010). As such, curricular standards emphasize the need for students to acquire and master the skills that are necessary to successfully learn from, use, integrate, and write about information represented across multiple documents (Achieve, 2013; Council of Chief State School Officers, 2010). Accordingly, there has been an increased level of research interest in the psychological processes and skills that support the use of multiple documents, as evidenced by this volume. The skills necessary to use and write about multiple documents are complex and involve challenges not often reflected in learning and writing about a single document (Anmarkrud, Bråten, & Strømsø, 2014; Rouet & Britt, 2011). To elaborate, imagine an academic situation where a student is asked to write an explanatory essay about a scientific process, which is not fully described in any one document at the student's disposal. Figure 1 represents such a situation. In order to write the essay, the student must identify information about the source, read each document, identify content relevant to the task within each document, extract that content, connect the relevant information extracted from each document to relevant information in the other documents, use that information to describe the process they are trying to explain, and identify the sources of their ideas in the essay. Doing these things can be particularly challenging when the documents are each written for a purpose that may not directly align with the task the student is using the document to complete (Magliano, McCrudden, Rouet, & Sabatini, in press).

[Figure 1 about here]

From the standpoint of a teacher or researcher, the essays produced by students are the external artifacts available to assess the extent that they were able to successfully engage in these processes. While tasks of this nature are challenging for students, evaluating and scoring essays and providing feedback are also complex and time consuming for teachers and researchers. Grading essays is always a daunting task and especially doing it such that feedback is provided in a timely fashion (Magliano & Graesser, 2012; Shermis & Burstein, 2013). There are a variety of dimensions that could be considered for assessment while scoring essays like spelling, grammar, cohesion, etc. (e.g., Magliano & Graesser, 2012), but in this chapter we are concerned about evaluating the content that reflects what was learned and used from a set of documents. In this context, one needs to evaluate how successful students (or participants) were at extracting information from the documents, synthesizing information, and indicating where their ideas came from.

The challenge of assessing essays for such tasks may deter instructors from assigning them to students. However, over the past two decades there have been substantial advances in the application of natural language processing (NLP) techniques to support the analyses of student essays (Magliano & Graesser, 2012; Graesser & McNamara, 2011). Generally, NLP refers to a wide range of computational approaches that are used to analyze the content, structure, and intended meaning of text. For example, computer programs have been developed to accurately identify the syntactic structure of sentences (Chen & Manning, 2014), the phrases in a paragraph which refer to the same objects (Clark & Manning, 2016), and the location of answers to questions (Morales, Premtoon, Avery, Felshin, & Katz, 2016). These advances can be brought to bear to create systems that are devoted to evaluating essays based on multiple documents (Hastings, Hughes, Magliano, Goldman, & Lawless, 2012; Hughes, Hastings, Britt, Wallace, & Blaum 2015; Wiley et al., 2017). We adopt a perspective that assessments in general should be grounded in theory (Mislevy, 1993; Pellegrino & Chudowsky, 2003), and in this case theories associated with understanding and learning from multiple documents (Rouet & Britt, 2011). As such, while we have emphasized the application of these tools in an educational context in this introduction, they could also be a boon for research on learning from multiple documents (Hastings et al., 2012; Higgs, 2016; Hughes et al., 2015; Wiley et al., 2017).

In this chapter, we first discuss features of essays based on multiple documents that are important to assess as delineated by theories of text comprehension and task-oriented reading (Rouet, 2006; Rouet & Britt, 2011). We focus on situations in which essays are based on a preselected set of documents rather than situations in which students find their own texts. To our knowledge, most existing systems were developed to address the former situation. Moreover, most research on learning and writing based on multiple documents reflects this situation (e.g., Anmarkrud et al, 2014; Blaum, Griffin, Wiley, & Britt, 2017; Wiley & Voss, 1999). We then discuss promising approaches to computer-based assessment of essays, and what is needed to develop and test systems specifically designed for essays based on multiple documents. We identify several challenges for developing these systems that are grounded in theory, research, and practical problems associated with automatic assessment of the use of multiple documents in essays. We present research on existing scoring systems that are illustrative of these approaches, but it is important to note that this area of research is in its early stages and there are only a few studies that involve the automatic scoring of essays. We conclude with a discussion of important directions for further development and testing of automatic grading systems for essays based on multiple documents.

Theoretical Perspectives On What Should Be Assessed

Evidence-centered approaches to academic assessment specify that these assessments should be grounded in relevant theories from cognitive science (Mislevy, 1993; Pellegrino &

Chudowsky, 2003; Pellegrino, Chudowsky, & Glaser, 2001). Specifically, theories of cognition associated with the academic activity should be used to identify constructs that are assessed. To this end, in this section we describe relevant theoretical perspectives of task-oriented reading (Rouet, 2006) and learning from multiple documents (Britt & Rouet, 2012). Based on these theories, we identify factors that should be assessed for essays based on multiple documents and some of the challenges for doing so when one is evaluating them with or without the aid of computational systems.

Theoretical constructs relevant to evaluating essays

In any reading situation, a person is reading in order to complete a task (Graesser, Singer, & Trabasso, 1994; Snow & The Rand Reading Study Group, 2002). Even people who are reading for pleasure are reading with the basic goal of understanding and hopefully enjoying the story or information they are reading. Purposeful reading has been described as *task-oriented reading* (McCrudden & Schraw, 2007; Rouet, 2006; Rouet & Britt, 2011; Vidal-Abarca, Mañá, & Gil, 2010). Task-oriented reading elicits goal-directed behaviors and strategies, which will vary based on the task the reader is trying to complete (McCrudden & Schraw, 2007). In Figure 1, the task orients the reader to the content from the documents for which they should allocate their attentional resources. Consider a situation in which the hypothetical task depicted in Figure 1 involves having to read a set of documents in order to generate a causal explanation for a physical process (e.g., Why are tsunamis destructive? How does coral bleaching occur? How can releasing carbon into the atmosphere lead to a rise in global temperature?).

Rouet and Britt (2011) provided a framework for how task-oriented reading may happen in a multiple-document reading situation, specifically the *Multiple Documents – Task-based Relevance and Content Extraction* (MD-TRACE) model. The MD-TRACE model describes how readers interpret a given task and create goals and strategies for completing that task. The goals and behaviors that readers will complete are part of their *task model*. As people read documents, they identify information that is relevant for the task they are trying to complete (McCrudden & Schraw, 2007). This information may be different than the information that is important to the underlying message of the document (McCrudden, Magliano, & Schraw, 2011; McCrudden & Schraw, 2007). As such, the overall message of a single document may not be relevant to the reader's task. Rather, some parts of that message may be used by the reader in order to complete the given task. Figure 1 represents this situation. In each document, only the information demarcated with a symbol is actually relevant for the student to complete the task of writing an explanatory essay based on the hypothetical prompt. Ideally, in this document set, students would be able to identify and extract the content germane to the task and discriminate it from that content which is not (Rouet & Britt, 2011).

While students are reading each individual document, they must be able to identify this information and begin constructing a *documents model* with information from multiple documents connected together (Rouet & Britt, 2011). After constructing the documents model, a reader will update the documents model regularly as they read through each document and encounter relevant information to complete their task. The situation reflected in Figure 1 involves constructing an explanation, and so the documents model would reflect a student's understanding of the explanatory processes. Finally, a reader will create a task product, the "Essay" shown in Figure 1, and then check whether that product satisfies their task goal. As students create their task product, they create and continually update a *product model*, which is a mental representation of what they have written. Initially, the product model is likely to have a high level of overlap with the task model, but as individuals write the document, they likely construct a mental representation of that document that is distinct from the task model. That is, the product model contains a representation of the content of the essay that likely is akin to a mental model for a text (Rouet & Britt, 2011). To assess whether or not their product has met their task goal, students will compare their product model with the goal created in their task model. This assessment is essential because as the writing process starts, the information that students decide to put into their writing product may deviate from what is necessary to meet their task goal.

Optimally, the task, documents, and product models should be conceptually linked, which is depicted in Figure 1. In Figure 1, no product model is shown because it is assumed that the essay is completed. As such, the essay is an external representation of the product model. If the task were to write an explanatory essay, one would need to build a mental representation of the information from the documents with regard to the goals related to the task. The task model would affect how information is selected, processed, and presented in the documents model (McCrudden, Magliano, & Schraw, 2011; McCrudden & Schraw, 2007; Rouet & Britt, 2011). Moreover, the task model should affect how content from the documents is integrated into a mental model within the document space. For example, if the task is to identify a causal explanation for a physical event (e.g., Why are tsunamis destructive?), the documents model would optimally contain a sequence of events that are connected via causal relationships (Griffin, Wiley, & Britt, 2016). However, if the task were to write an argumentative essay about some issue (e.g., Write an argument about various preventative steps that should be taken to lessen damage from tsunamis), then the documents model will be structured around claims, reasons, and potentially counter-arguments. The essay that is produced by students should be indicative of the task, documents, and product models (Rouet, 2006).

Dimensions of essays that should be evaluated

This chapter emphasizes evaluation that pertains to the use of the documents in a set provided to the student. First, it is important to assess the extent to which content was used from the different documents. However, students could summarize the relevant content, but do so in a manner that does not reflect the goal of the essay (e.g., describe a causal process, compare and contrast positions about an issue, write an argumentative essay that supports a particular position). As such, a second issue to consider is whether the content is conveyed in a manner that reflects the task at hand. An ideal essay should reflect the relevant content in a manner that is consistent with the prompt (e.g., the explanation of a process), rather than being constrained by how the content was conveyed in the documents.

A third issue to consider is *sourcing* (Britt & Aglinskas, 2002; Rouet & Britt, 2011; Wineburg, 1991). Sourcing refers to activities that involve evaluating the reliability of the sources, and documenting how the sources were used in the essay (Wineburg, 1991). Monitoring source reliability is important because, although the information may be relevant to the task, that information may not be true. Monitoring source reliability involves assessing the author's expertise on the subject, the outlet where it is published, the intent of the author, and possibly the date of the publication (Bråten, Strømsø, & Britt, 2009; Britt & Aglinskas, 2002; Wiley et al., 2009; Wineburg, 1991). This is important when, for example, there may be conflicting information from two sources. In such cases, it would be better to use information from a trustworthy source.

Unfortunately, most people are not overly sensitive to the author (Britt & Aglinskas, 2002; Claassen, 2012) or vigilant about keeping track of the authors when processing a set of documents without training (Bråten et al., 2009; Britt & Rouet, 2012). Nonetheless, teachers and researchers should be sensitive to whether or not students are drawing upon the sources in a document set appropriately, sufficiently transforming that content to meet the task and avoid plagiarism, and following protocols for indicating where the ideas came from.

What Is Needed to Evaluate Essays Based On Multiple Texts

We are considering a situation in which there is a defined document set given to students or participants that is the basis for their essay (e.g., Blaum et al., 2017; Wiley & Voss, 1999). This is different than a situation in which students self select texts that are unknown to an instructor. We are restricting this discussion to a closed-documents-set context because the automatic assessment protocols that have been developed to date reflect that situation. Based on the discussion above, we identify what an instructor or researcher needs to implement a writing task based on multiple documents. We describe these here because they are also germane to developing systems that automatically evaluate essays.

1. **An essay prompt.** Prompts should ideally be specific enough to support the development of a specific task model that allows the reader to process the texts in a strategic fashion. Moreover, these prompts should clearly state the instructor's desired structure for the essay, whether it is a causal prompt, argumentative prompt, compare and contrast, etc.
2. **A documents set**. Texts need to be selected that contain content that can be used to write the essay. A decision needs to be made as to the extent that the purposes of the documents are aligned with the essay prompt. The more divergent the intentions of

the authors of the documents are from the essay prompt (i.e., the texts were written to convey points different from how their contents should be used in the essay), the more challenging it is for the student to identify relevant content from the documents. The more divergent the documents are from each other, the more difficult it is for the student to integrate information between documents, but the easier it is to identify where the information came from.

3. **A documents model**. A representation that contains an idealized specification of 1) the content from the documents that is relevant to the prompt, 2) how that content should be linked together to address the prompt (e.g., bridging inferences that explicitly link ideas in a manner consistent with the prompt), 3) how that information is transformed and synthesized across multiple documents (i.e., ways in which students may transform ideas in the texts to address the prompt), and 4) where that information came from.
4. **A scoring rubric**. A set of dimensions used to objectively evaluate the contents of a product model and how well it satisfies the task given by the essay prompt.
5. **A protocol for delivering feedback**. Feedback that is provided should be timely, appropriate, and targeted to address any deficiencies detected by the scoring rubric. The purpose of the feedback should be to help individuals modify and improve their task and product models, and therefore, ultimately the quality of their essays.

This list reflects not only what is needed when one develops and implements a writing task that involves multiple documents that will be evaluated by an instructor or researcher, but also situations that would involve automated systems. These five dimensions have important challenges to overcome. However, before we discuss these challenges, it is important to understand how prevalent approaches to the automatic assessment of student constructed responses work.

Promising Approaches in Automated Essay Assessment

There are many different natural language processing techniques that have been developed for a wide range of purposes. In this section we describe the major categories of approaches and present some of the more prominent specific techniques within those categories, including the supporting resources that are required to use them. We then describe some example applications of their use in the evaluation of multiple source use in essays (Hastings et al., 2012; Higgs, 2016; Hughes et al., 2015; Wiley et al., 2017).

General overview of scoring systems

There are several core features of any system that is designed to assess multiple document use, of which most correspond to the five features that are necessary for a multiple documents essay task that are described above. Some features are a requirement of all systems, whereas other features constitute options for the developers. In this section, we discuss some of these features.

The first key features are s*emantic benchmarks,* which are critical for assessing content that should be in an essay as delineated by the documents model**.** Many computational systems that evaluate constructed responses (e.g., essays, answers to open ended questions, think aloud protocols) do so by comparing those responses to *semantic benchmarks* (Magliano & Graesser, 2012). Semantic benchmarks can reflect a variety of things, such as idealized responses (e.g., Magliano et al., 2011), a set of responses that vary in quality (Foltz, Gilliam, & Kendall, 2000), or content specified by theory or an ideal model as being important for the response (e.g., Magliano & Millis, 2003; Magliano et al., 2011; Hastings et al., 2012). If the system is designed to detect the extent that students identify the sources of their ideas, then semantic benchmarks could be created that reflect how the students are instructed to do this (e.g., list authors' names). The underlying assumption is that the more essays have semantic overlap with the semantic benchmarks, the more those benchmarks are reflected in the essays. As an example, consider the situation reflected in Figure 1. Three semantic benchmarks could be constructed for the relevant content from the three texts that should be included in the documents model (i.e., benchmarks corresponding to the bolded symbols in each document box). Essays that would have a high level of semantic overlap with all three semantic benchmarks would be considered more compliant with the prompt than those that have low overlap with one or more benchmarks.

A second feature is the grain size of the text and benchmarks. In order to compare an essay to semantic benchmarks, one must determine the grain size of what is being compared. The grain size can vary from the whole document, to groups of sentences, to single sentences, to individual words or phrases. Comparing an essay to semantic benchmarks that reflect entire texts allows a general assessment of the extent that the contents of the texts are manifested in the essay, but does not allow one to assess the extent that specific content is included. On the other hand, smaller grain sizes may allow for some level of precision in determining what specific content from a text is actually manifested in the essay, but might not provide a good holistic evaluation. Importantly, parsing essays into smaller units of analysis is essential for determining if content across a document set is integrated into the essay. Consider Figure 1, which depicts a situation that requires students to integrate content from the three documents. As such, an essay that would contain the content, but in a compartmentalized fashion (e.g., text 1 content is described, then text 2, and finally text 3, rather than as is depicted in the explanatory process) would not be compliant with the prompt. However, an assessment system based on text-level grain size would not be able to distinguish essays that are compartmentalized from those that demonstrate integration. We contend that successful systems for scoring essays based on multiple, pre-designated documents require a grain size of comparison smaller than the essay or texts in the documents set.

A third feature that is essential for system development is a corpus of human-scored essays based on a closed set of prompts and document sets, where the scoring is based on a rubric (as discussed above). That is, the corpus should reflect the exact prompt(s) and document set(s) that will be used once the system is developed. The essays must be scored on the dimensions that the system will be designed to evaluate. There is no specific number or type of features that should be scored, but ideally they should be delineated by theory (Mislevy, 1993; Pellegrino & Chudowsky, 2003; Pellegrino et al., 2001). According to the MD-TRACE model (Rouet & Britt, 2011), the scoring system could be constrained by 1) evidence of sensitivity to the essay prompt (i.e., the task model), 2) evidence that it reflects the ideal documents model (both content and how that content should be integrated), and 3) the extent that students have identified the sources of their ideas. That said, the systems that have been developed to date (and are described below) have primarily scored the essays on the extent that they reflected integrated content, rather than being sensitive to the essay prompt (beyond overlap with the documents model) and explicit sourcing.

Another potential aspect of human scoring is the annotation of the essays such that specific content is linked to idea units (clauses, sentences, groups of sentences, etc. (Hastings et al., 2012; Hughes et al., 2015; Wiley et al., 2017). Consider Figure 1. While not always necessary for scoring systems (for example, where the goal is not to determine coverage of the sources by the essays), essays could be annotated such that content in the essays is specified as being semantically connected to sentences in the texts that are identified as being important to the documents model.

Annotation is particularly valuable when machine-learning algorithms are used to train a system to identify if essays contain content that is aligned with the documents model. As such, a fourth feature of scoring systems is training based on machine learning. Machine learning refers to computer algorithms that automatically learn from data (Mitchell, 1997; Russell & Norvig, 2010). Annotated essays and documents are required for training because they provide the data used to train the system. Specifically, the system learns to classify a variety of ways that content from the text can be presented by students. For example, consider Table 1, which shows two idea units from a document set on coral reefs and a sample of protocols that reflect how participants described those ideas unites in the context of writing an essay on the process of coral bleaching (Kopp et al., 2016). Protocols 1a and 1b reflect a close paraphrase, whereas protocols 5a and 5b reflect some degree of transformation of the original text content. The more an idea from a text is transformed by the students, the more challenging it is to computationally determine the source of that idea. As such, creating a semantic benchmark that reflects the variety of ways that a sample of students can express relevant content from the text can be useful, and is essential in any system that is trained to detect the variety of ways ideas can be expressed. This is an issue that is of concern for the computational analysis of student-constructed responses in general, whether they be essays (e.g., Hastings et al., 2012), think aloud/self-explanation protocols (Millis, Magliano, Todaro, & McNamara, 2007), or question answer protocols. While some systems for analyzing student-constructed responses do not require the system to be trained to detect overlap between essays and semantic benchmarks (e.g., Magliano et al., 2011), some systems have relied on machine learning to do so (e.g., Hastings et al., 2012; Millis et al., 2007).

Below we will describe two studies that have explored the extent to which training benefits essay scoring in the context of multiple documents (Hastings et al., 2012; Hughes et al., 2015).

The final feature of scoring systems that we highlight is scoring categories. Developers have to determine exactly what aspects will be scored, the data for evaluating those aspects, an operationalization of the dimensions (i.e., the scores), and if feedback is provided to users (teachers, students), protocols for delivering those scores. There are no hard and fast rules on developing scoring criteria, as they are ideally constrained by theory, research questions, and or practical considerations (e.g., curricular decisions, constraints on what can be assessed). For example, the Reading Strategy Assessment Tool (RSAT) developed by Magliano and Millis (Magliano et al., 2011) was designed to score think aloud protocols. They identified two types of inferences (i.e., bridging and elaborative inferences) as a scoring dimension because they were delineated by theory to be important for comprehension. Scoring dimensions for analyzing essays based on multiple documents should be dovetailed with the dimensions that were identified as being important for scoring the essays by human coders in the scoring rubric. While we mention the importance of developing feedback for the user, the nature of that feedback depends on the user. To date, the systems that have been developed to score essays based on multiple documents have been developed for research purposes and therefore protocols for delivering feedback to teachers and students have not been developed.

Computational tools for analyzing essay content

In this section, we describe techniques for evaluating the content of the essays, parsing essays, and training systems to evaluate essays for specific content.

Approaches for analyzing semantic content. There are two general approaches for comparing student responses to semantic benchmarks, which is typically in the service of identifying the content of an essay, but could also be used to identify the extent that students explicitly identify the sources of their ideas. The first is *keyword matching and regular expressions* (Magliano & Graesser, 2012; Hastings et al., 2012). The simplest indication that something in an essay was derived from a particular source is that it uses the same unique words to describe whatever that is. By "unique" we are referring to words that occur in one of the given sources, but not in the others. Simple scanning techniques can search essays for important terms (keywords) or consecutive words (known as *n-grams*) that comprise the semantic benchmarks created for scoring the essays. If the benchmarks reflect the different sources, this search can be used to identify the relevant source.

Ideally, however, students will transform the content of the essays because they are instructed to convey it in their own words. As such, there can be a "family" of ways in which content could be expressed by students (See Table 1). An alternative to keyword matching is to identify a set of expressions or patterns that might reflect the different ways in which semantic benchmarks can be conveyed. These are normally referred to as *regular expressions* (Aho, 1990). Regular expressions_provide a way of specifying keyword strings that include variants. For example, the expression, "increas(ingled) fresh water" can match 2 different key phrases, "increasing fresh water" or "increased fresh water." By combining regular expressions, one can specify key phrases in a way that is rich, powerful, and concise. All modern programming and scripting languages include built-in mechanisms or libraries for searching for regular expressions. (Keywords can be treated as simple regular expressions.)

A second approach for evaluating essays against semantic benchmarks involves the use of a more general (i.e., not customized to a particular task) high dimensional vector or semantic space (Magliano & Graesser, 2012), such as Latent Semantic Analysis (LSA; Landauer &

Dumais, 1997), Hyperspace Analogue to Language (HAL: Lund, Burgess, & Atchley, 1995), holographic models (Jones, Kintsch, & Mewhort, 2006), and word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Because these techniques create vector representations of words and/or documents, they are commonly referred to as Vector Space Models (VSM) of Semantics (Turney & Pantel, 2010). A number of these VSM models ignore word order, in which case they are referred to as *bag-of-words* models. All of these models are based on the Distributional Hypothesis of word meanings, which holds that words which occur in similar contexts tend to have similar meanings (Firth, 1957).

Approaches like LSA and HAL start with the creation of a co-occurrence matrix that reflects the extent that words co-occur across a large set of (possibly domain-specific) texts. (The web site http://lsa.colorado.edu contains a number of previously-developed spaces reflecting different topics and ranges of texts.) The matrix generally contains thousands of words and the frequency at which they co-occur across thousands (or more) of texts. With LSA, a dimensionality reduction technique, Singular Value Decomposition, is used to reduce the number of dimensions from thousands to typically 100-500. Word “meanings” are represented as vectors within the semantic space. Similarity of words can be simply computed by calculating the proximity of the vectors of the words, typically using the geometric cosine, which, in practice, varies from 1.0 (semantically identical) to near 0 (completely unrelated). For example, using the Colorado LSA space representing general reading up to the first year of college, “tsunami” and “wave” have a cosine of .76, whereas “tsunami” and “mountain” have a cosine of -.01, indicating that the first pair are semantically very close in the semantic space and the second pair is essentially unrelated. Representations for groups of words (sentences, paragraphs, texts) are computed by a simple combination of the vectors for the words in the groups. A semantic space approach can thus be used in a multi-document setting by using it to compare the sentence (or words or paragraphs) of a new text to the original source documents and/or semantic benchmarks to identify those with a sufficiently close “fit” (typically using an empirically-determined cosine threshold).

One advantage of the semantic space approach over keyword matching is that it is sensitive to semantic distance (as reflected in the example from the Colorado LSA space above) and therefore does not require one to develop dictionaries of synonyms or a family of regular expressions. However, the use of keyword matching, regular expressions, and high dimensional spaces to analyze constructed responses are not mutually exclusive, and there are examples of hybrid systems (e.g., Graesser et al., 2004). In fact, it has been argued that systems should rely on both approaches to compare constructed responses to semantic benchmarks whenever possible (Magliano & Graesser, 2012).

Approaches for parsing sentences in essays. Semantic space methods can be used at different levels of granularity. They can be used to compare entire essays, or paragraphs, or sentences, or words. But they do not provide information about the relationships between, for example, the words in a sentence. There are situations in which parsing essays is advantageous in the computer-based assessment of essays. Parsing involves determining the structure of the sequence of words in sentences and the phrases within them. For sentences, one may apply a syntactic grammar (often along with semantic constraints) to determine the phrasal structure. The structure is normally viewed as a tree, normally with the main verb as its root, and the phrasal attachments as the branches. From the syntactic structure, the specific semantic relationships between the components of a sentence can be derived. This type of analysis is often necessary in order to gain a clearer understanding of the meaning of a text. For example, in the sentence, “The woman kissed the man," a bag-of-words semantic space approach would not be able to determine who is doing the kissing because the order of the words is not evaluated.

Both semantic space methods (when applied at the sentence level) and parsing-based methods benefit from the segmentation of a text into sentences. This is normally easy to accomplish, especially with the use of punctuation (assuming it is reliable and assuming exceptions like abbreviations are taken into account). If the semantic benchmarks for a task are relatively specific (as is depicted in Figure 1), the system may be more accurate in detecting them if the essay text is parsed, because then it can compare concepts from the benchmarks to representations of specific phrases. Parsing is also necessary when the different documents in a set describe the same entities, but with different relationships between them. For example, in the document set used in Hastings et al. (2012), one document described how the advent of the trains allowed Chicago to become a transportation hub, whereas another described how trains made it easy for people to move to Chicago. Discriminating these different roles of trains in Chicago population growth might be improved by automatically analyzing the clauses in the student essays that describe trains. The presence of nouns that are strongly associated with the concept "trains" (e.g., train, locomotive, tracks, etc.) provide semantic cues that these sentences are associated with that concept. However, the verbs associated with the roles of trains should be indicative of the events associated with them from the documents that the students are describing (e.g., Zwaan, Langston, & Graesser, 1995). These verbs can provide cues to causal and situational cohesion (McNamara, Graesser, McCarthy, & Cai, 2014), and therefore could be useful in detecting the extent that students are linking ideas in the essay in a manner consistent with the documents model.

As noted above, many semantic space systems do not consider word order, and one reason that semantic space methods are so popular is that parsing can be a notoriously difficult task. Besides ungrammaticalities or partial phrases introduced by the writer, the biggest problems are due to the inherent ambiguities of human languages. Many, if not most, words have multiple senses, phrases and discourse elements can be combined in different ways, and ambiguities at lower levels multiply the ambiguity at higher levels. For example, if a word has two possible interpretations, and that word is in a prepositional phrase that can attach to another element in the sentence in three different ways, that leads to six different interpretations of the sentence. This ambiguity can make parsing computationally very intensive and make it very difficult to determine the intended meaning.

Recent research into this problem has reduced it substantially, by using probabilistic grammars that are learned from real-world texts (e.g., Chen & Manning, 2014). These grammars take into account the likelihood of the various combinations, and only pursue the most likely. The existence of very large annotated corpora has also allowed these grammars to include semi-semantic information like dependencies, which go beyond the basic syntactic relationships between words. Dependencies indicate which word in each phrase is the root and the types of relationships between the roots and the other words. For example, “advmod” indicates that the word is an adverbial modifier to a root verb, an “amod” is an adjectival modifier, and an “agent” dependency indicates the performer of a verb’s action (de Marneffe et al., 2013). The best known of these systems is the Stanford Natural Language Processing Group's CoreNLP system (available from http://nlp.stanford.edu/software/).

Custom Machine Learning methods. Some types of machine learning require no special annotation (i.e., where the researcher designates the meaning of semantic units in a training corpus). This is called *unsupervised learning*. In this section, however, we focus on *supervised learning* approaches to assessing multiple documents use, where the training data has been coded by human coders to indicate whether it does or does not fall into particular categories.

For example, assume a task for which researchers have developed a causal model, which includes all of the main concepts and the desired causal relationships between them. (The next sextion provides a more detailed description of such a situation.) The training data would consist of a large number of example (student) texts (at least two hundred, preferably many more) in which human coders have identified specifically where the different concepts (from different sources) and connections between them are mentioned. For this task, it is useful to use an annotation tool like brat (available from http://brat.nlplab.org/index.html). Then a Machine Learning algorithm such as Support Vector Machines, Logistic Regression, or Neural Networks can be used to infer from the examples how to identify the concepts and relationships in a new text (Mitchell, 1997; Russell & Norvig, 2010).

Example studies using NLP to study multiple documents processing

In this section we discuss four studies that have used NLP systems to study multiple documents processing.

Using NLP tools to identify source content. Hastings et al. (2012) conducted a study to test different methods of automatically assessing the content of essays written from multiple documents, specifically: LSA, identification of keyword phrases (n-grams) with machine learning, and the machine learning technique called Support Vector Machines (SVMs), which learns the best separation of texts into classes (here, whether or not they contain specific concepts from the ideal products model) based on the words that occur in them. In their study, 460 essays were collected from students in grades 5 through 8. These students were asked to use three text-based documents to answer a question about why people moved to Chicago between 1830 and 1930. A documents model was created to identify sentences from the document set that were relevant and how they should be integrated in the essay to answer this question. Human coders segmented the student essays into sentences and then identified the extent to which each sentence reflected the ideas specified in the documents model.

In the LSA approach, the semantic benchmarks were simply the human annotations that reflected which sentences in the document set corresponded to the parts of the idealized product model. LSA cosines were computed between each sentence in the essays and the semantic benchmarks and a threshold was determined to indicate if the sentences were similar enough (the cosine was high enough) for a model concept to be considered present in the essay. The Machine Learning approaches used ten-fold cross-validation, learning from a randomly-chosen 90% of the essays, and testing on the other 10%, averaging performance over 10 iterations.

Performance of the systems was calculated by comparing their identification of concepts from the products model to those of the human annotators. Overall, the LSA approach most closely matched the human judgments over the entire set of concepts when the frequency of the concepts was taken into account. However, some of the concepts relied on connections between sentences in a single document or across documents. In these cases (especially the latter), the LSA approach performed poorly because it was based on sentence-to-sentence comparisons. The overall performance of SVM was similar to that of the LSA approach, but it suffered when concepts occurred rarely in the corpus of essays. This is a typical problem for Machine Learning approaches; the frequency of the answer in the training set affects the frequency that the answer is given during testing. For detecting content that made connections between documents, the ngram learning method performed best. It was trained on each concept individually, so it was unaffected by the frequency of occurrence of the concept, and, unlike the LSA approach, it did not rely on matching specific sentences in the source documents.

Hughes et al. (2015) developed an automatic coding system to assess the overall quality of causal essays written from multiple documents based on the content and structure of the essay. Students were given a set of five documents about the topic of coral bleaching or skin cancer (each set had 4 text documents and 1 graph) and asked to write about what causes the scientific phenomenon they read about. Human coders scored and tagged the essays for key content and the causal links made connecting the content. Based on the amount of content and number of connections students made, their essays were sorted into four different quality categories: no core content from the documents, some core content, but no connections between any content, some core content, but only one connection linking content, or both content and at least two consecutive connections structurally linking the content. This task is especially difficult because any misclassification on the component tasks (identifying concepts and causal connections) affects performance on the holistic assessment. A machine learning system was trained on a subset of essays which had been tagged by human coders. Once trained the system could place an essay in its appropriate quality category with moderate success, with a Krippendorff's alpha correspondence with human coders of 0.56 for essays about coral bleaching and 0.47 for essays about skin cancer. If the neighboring quality category was included (indicating that the system was not far off in its assessment), then the accuracy was 85% and 88% respectively for the two topics.

Using NLP tools to assess explanation quality and student understanding. Wiley et al. (2017) built on the machine learning-based research described above and compared it to other methods of evaluating multiple source use with the goal of assessing how well each method accounts for both student understanding of the information from multiple documents and the quality of the explanatory essays that they created. The participants in this study were 178 middle and high school students. The students were given 7 short documents on topics related to global temperature change. One document gave general background information, 5 described related main topics like ice ages and the carbon cycle, and one was a graph of carbon dioxide concentrations over the last 400,000 years.

The students were asked to read the documents and then write an essay (with the documents present) to explain "how and why recent patterns in global temperature are different from what has been observed in the past." None of the source documents was sufficient by itself to create a complete answer to the prompt. The essays were annotated by human scorers as described above, and quality categories were derived from the coding of concepts and connections. After writing the essay, the students were given an 18-question multiple choice inference verification test to assess their understanding as indicated by the connections and inferences that they made within and between the documents.

Along with the machine learning method described above (Hughes et al., 2015), this study analyzed metrics derived from two "off-the-shelf" tools. LSA was used in two ways. Following Ventura et al. (2004), an ideal essay similarity score was computed by comparing the student's entire essay to an idealized essay that was constructed from 2 highly-rated peer essays. Following Britt, Wiemer-Hastings, Larson, and Perfetti (2004), plagiarism scores were calculated by comparing each student essay sentence to each sentence of the source documents. If the maximum cosine was above 0.75, the sentence was deemed to be copied from the source.

The plagiarism score for an essay was the percentage of its sentences that were marked as copied.

Coh-Metrix was also used, which is an online tool (available at http://cohmetrix.com) providing 108 indices of a variety of aspects of readability, cohesion, and complexity of texts (McNamara et al., 2014). From these, three metrics were computed: causality, cohesion, and lexical diversity. The causality score operationally was based the number of causal verbs and particles. The cohesion score was based on LSA cosines between paragraphs in the essays. If there was only one paragraph in an essay, cosines between adjacent sentences were computed instead. The lexical diversity score was the type-to-token ratio of the essay, which is the number of unique content words which appeared in the essay divided by the number of occurrences of those words.

These metrics, along with some basic descriptive features of the essays were entered into simultaneous regression equations to see how well they predicted the overall essay quality scores and student understanding of the documents as indicated by the inference verification test. For essay quality, the unique significant predictors of variance, predicting a combined 49% of the variance, were the number of concepts identified by the machine learning model, the LSA plagiarism score, the LSA comparison to the idealized essay, and the Coh-Metrix cohesion score. The unique significant predictors of variance in student understanding, predicting a combined 23% of the variance, were the number of concepts identified by the machine learning model, the LSA ideal essay similarity score, and the Coh-Metrix causality, cohesion, and lexical diversity scores.

This study has a number of interesting conclusions and implications for future research. One is that automatic methods of assessing explanatory essay quality are feasible, but that is especially so with hybrid models that combine a number of different types of factors (Magliano & Graesser, 2012). Surprisingly, the basic text features of essay length, responsiveness to the prompt, and presence of citations were not found to be related to essay quality. This has implications for studying writing processes; the lack of predictive power of citations could indicate that students who referred more to the source documents were focusing more on knowledge-telling rather than knowledge-transforming (Wiley & Voss, 1999). It also suggests that more research is needed on the machine learning approach to identifying connections between concepts. Current approaches are limited in the extent to which they take into account discourse features like anaphora which can play a large role in explanations.

Using NLP tools to study how texts are processed in a multiple documents task. While the emphasis of this chapter is on the development of systems for analyzing essays based on multiple documents, NLP tools can also be used to study how texts are processed in a multiple documents task. For example, Higgs (2016) was interested in assessing if integration across documents happens during reading or after reading (i.e., while engaged in the writing task). She manipulated the specificity of a reading goal that was either general or emphasized a specific topic in the texts. Participants first read the texts silently under the three instructions (read to understand, read to learn about tsunamis, and read to explain why tsunamis can be destructive). None of the texts were specifically about tsunamis and why they can be destructive, but an explanation for that could be derived from content across the three texts. They then re-read the texts and thought aloud at target locations. A documents model was created by the experimenters that reflected how content across the texts could be used to answer a causal prompt about the text (i.e., *Why are tsunamis so destructive*?). Sentences were selected for the think aloud prompts that afforded integration of content in the documents model. These sentences were chosen because readers should make connections to other sentences in the same text and across texts in the document set.

Inspired by studies that have used NLP tools to analyze think aloud protocols (Magliano & Millis, 2003; Magliano et al., 2011), Higgs used LSA to analyze the think aloud protocols to assess if they reflected intra-text (i.e., within text) or inter-text (i.e., across texts) integration. Specifically, she compared the verbal protocols to two semantic benchmarks: one that reflected the most important ideas in the texts with respect to their topic (which was not specific to the causal prompt) and another benchmark that reflected content from the three texts that was in the documents model. Higgs found that cosines were higher for the important ideas benchmark than the documents model benchmark under the general comprehension instruction, but were the same under the two specific reading goal instructions. She concluded that specific task instructions led readers to process content semantically aligned with the prompt more closely than the general instruction (e.g., McCrudden & Schraw, 2007).

A second analysis involved developing two benchmarks associated with content in the documents model. Specifically, an intra-text benchmark was constructed that reflected the content in the documents model that was in the text that was being read when the verbal protocols were produced. An inter-text benchmark reflected content in the documents model from other texts (i.e., not the text that was currently being read when the protocols were produced). Higgs found that under the goal to explain why tsunamis are destructive, the LSA cosines were higher for the intra-text benchmark than the inter-text benchmark, but there were no differences under the other two instructions (and the cosines were relatively small). She concluded that under read to explain instructions, participants focused on making connections to content aligned with those instructions in the text that was currently being read, but did not make connections to other documents in the set. Importantly, specific task instructions (both to learn about tsunamis and read to explain) did lead to evidence of greater inter-text integration in the recall protocols (as evidenced by human coding of the protocols) than the general instruction. Higgs (2016) concluded that integration across documents likely occurred during the writing task rather than during reading.

Challenges in developing automated essay evaluation systems

In the first section of this chapter, we discussed challenges instructors and researchers face when grading and evaluating essays based on multiple documents. The goal of this section is to discuss how these challenges need to be addressed to develop a system that can be used to automatically analyze essays based on multiple documents. For some of these problems there are good solutions, and other problems have not yet been solved.

Level of semantic overlap between documents in a document set. The level of semantic overlap between documents in a document set can vary. For example, the documents used in Higgs (2016) were all written on different topics (i.e., the role plate tectonic shifts play in earthquakes, how tsunamis form, and the Fukushima Nuclear power plant disaster), but each provided part of the causal process specified by the question prompt (Explain how tsunamis can be destructive). On the other hand, the documents used in Hastings et al. (2012) each explained one aspect of how Chicago grew in population during the later half of the 19th century (i.e., the availability of jobs in Chicago, economic problems in the south, Chicago becomes a transportation and shipping hub). As such, there may be stronger cues to facilitate integration within the documents used in Hastings et al. (2012) than Higgs (2016).

There is a tradeoff one needs to consider in terms of having documents with sufficient semantic overlap to support integration when students are writing their essays but with sufficient semantic dissimilarity to be able to computationally detect which documents are being used in the essays (Hastings et al., 2012). It is easier to integrate content across documents in an essay when those documents describe similar events than when they describe different events (Kurby, Britt, & Magliano, 2005). On the other hand, the more semantic overlap that there is between documents, the harder it is to develop a system that can accurately detect the extent that content from the different documents that is specified in the documents model is present in the essays. This can be particularly challenging when the documents include graphs to be interpreted by students. In Table 1, protocol 5a represents the idea present in the given idea unit from the text, but the document set also included a line graph which included temperature changes over time. While it is possible, and even likely, that the student used the word "spike" to indicate the visual cue on the graph, it is difficult to tell for certain whether the student was transforming the idea from the text or using information present in the graph on a separate document.

Research is needed to specify the right balance between overlap that affords integration and maximizes source detection. However, a first question should be, "How well can human judges distinguish the source of the concepts that the students write about?" With enough examples, modern computational methods are very good at distinguishing sources. But they cannot perform well if their training data (annotated essays) are unreliable.

Student transformation of content from the documents. As discussed above, ideally students should transform the content from the documents such that it is described in their own words. However, given the propensity of students to closely paraphrase source documents and even "write" by cutting and pasting, a number of essay grading systems have developed protocols to detect plagiarism (Foltz et al., 2000). However, in this chapter we want to emphasize the challenge of students transforming content into their own words. For example, consider

Table 1. It would obviously be much easier to determine the semantic overlap of the source text with Protocol 1a rather than with Protocol 5a. Both protocols are paraphrases of the source texts in that they convey the same idea (temperature increases). Protocol 1a contains many of the words in the source text, but Protocol 5a does not. As such, cosines based on the Colorado LSA space are very different.

This is exactly the situation in which the development of regular expressions and machine learning can be used to train a system to recognize the different ways students can produce content from a document set. For instance, from an annotated corpus, a machine learning model could detect features that are commonly associated with expressing a certain concept in text. In addition, unsupervised machine learning approaches to modeling semantics, such as LSA and word2vec, can make use of semantic information acquired from large external corpora to map the annotated essay text into a semantic space, and then supervised machine learning models can be trained on this representation. This allows the system to recognize other ways of expressing these ideas that were not observed in the annotated corpus.

Dedicated versus general semantic spaces. If one is using high dimensional spaces, such as LSA, an important consideration is whether one needs to develop a dedicated semantic space that covers the topics in a document set. Many systems that have been developed to code student-constructed responses have relied on general spaces, such as the Colorado TASA (Touchstone Applied Science Associates) LSA spaces (e.g., Foltz et al., 2000.; Hastings et al., 2012; Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007; Magliano & Millis, 2003), which were built from corpora including representative texts that might have been seen up to 3rd, 6th, 9th, and 12th grades, or college. However, if the documents in a set describe relatively novel or specialized topics, then those specialized word senses might not be represented in the general semantic space. For example, Higgs used texts describing geological processes associated with tsunamis (Higgs, 2016). While she used LSA to analyze the think aloud protocols, many of the words in the documents were not represented in the semantic space (e.g., subduction). As such, LSA was not sensitive to their presence.

There are two options present to the developers. The first is to develop a dedicated high dimensional space (Kurby et al., 2003). This is a time consuming process that not only involves developing the space (i.e., collecting a large amount of domain-specific text and creating the space from it), but testing its validity. A second option is to rely on a hybrid system that uses both semantic spaces and keyword matching/regular expressions. Given the time-consuming nature of developing and testing a new semantic space, we advocate using hybrid systems (Magliano & Graesser, 2012).

Detecting the relationships between content. Documents models specify both content that should be in the essays and the relationships between them. While we have developed protocols for assessing the presence of specific content, as discussed above, there are significant challenges to assessing the relationships between content, as specified in the documents model. Often students may explicitly state the idea units in their essays, but they do not always explicitly state the relationship between the ideas. Without explicit markers, it can be quite challenging for an automated system to detect these relationships. Additionally, students will connect ideas across sentences. For example, a student might write, "Sometimes weather changes and trade winds decrease. This causes ocean temperatures to increase." It is challenging to determine exactly the antecedent for the referent "this." A human may be able to intuit that it refers to the decrease of trade winds by applying her general world knowledge, but a computer doesn't have that luxury.

Developing systems that can generalize. To date, most systems that have been developed to analyze essays based on multiple documents have been specialized systems that are specific to a document set and essay prompt. While it is possible to develop a system that can handle different essay prompts for the same document set, there are significant challenges to developing a system that is generalized enough to handle any document set.

Challenges to providing feedback. None of the systems that have been developed to analyze essays based on multiple documents are designed to provide feedback to users (students or teachers). As such, this has not been a primary focus of this chapter. Nonetheless, we point this out as a challenge to overcome. Of course, there are systems that provide feedback to users about their writing in other contexts (e.g., Dai, Raine, Roscoe, Cai, & McNamara, 2011; Kintsch et al., 2007). The nature of that feedback is determined by the pedagogical goals. Given that those goals can vary, we do not belabor the point here. Rather, our goal is to acknowledge that this is a dimension that developers need to consider.

Conclusions and Future Directions

This chapter is best seen as a primer for developing systems that can support the automatic assessment of essays which are based on multiple documents rather than a chapter that specifies the technical features of these systems. The systems that have been discussed in this chapter pertain to the evaluation of essays based on a closed set of documents with an emphasis on evaluating document use. We have discussed the features of such tasks (i.e., document sets, prompts, documents models, scoring rubrics), dimensions of systems designed to evaluate essays (i.e., semantic benchmarks, specified grain size of content that is assessed, human scoring, and training), promising approaches for addressing these features (keyword matching and regular expressions, high dimensional semantic spaces, automatic parsing, and machine learning), and finally challenges for implementing systems (optimizing overlap between documents, evaluating transformed content, developing dedicated semantic spaces, creating systems that can be generalized, and providing feedback).

To date, we know of only two systems that have been designed to evaluate essays based on multiple documents (Hastings et al., 2012, Hughes et al., 2015), and as such, more research is needed to learn how to best develop these systems. Based on what we have learned in developing these systems, and the challenges raised in the last section, we conclude by identifying key areas that warrant more research to support the development of automatic scoring systems for multiple document use.

One pressing issue discussed above is learning how to optimize the level of semantic overlap between documents to maximize the ability to detect their use and provide sufficient semantic scaffolds to afford using those documents to address the task specified in the prompt. The insights gained into this issue arose when developing systems for studies in which this issue was not under consideration (Hastings et al., 2012; Hughes et al., 2015; Wiley et al., 2017). Given that integration is facilitated by the extent that documents in a set discuss similar events (e.g., Kurby et al., 2005), we envision studies that systematically manipulate the level of semantic overlap at the level of events. Criteria would need to be determined for evaluating both the success at which students are able to integrate texts and the success of the automatic evaluation of document use.

A second pressing issue is developing approaches to determine the extent that content reflects the relationships specified in the documents model (e.g., causal, logical, and argumentative). Evaluating the extent that explicit content from the documents is in the essay is a relatively less challenging problem because there are explicit semantic relationships that can be assessed between the essay and documents (i.e., content from the essays can be compared to semantic benchmarks reflecting content from the texts). It is possible to develop a system that can determine explicit linguistic markers of semantic relationships, such as the use of appropriate connectives (logical, temporal, and causal connectives) between important idea units in the essays. However, students may leave these out and rely on conveying the relationships implicitly. One possible solution was described above in the context of Hughes et al. (2015). They developed an assessment protocol to detect relationships using machine learning algorithms. They relied on the gold standard of human judges to classify essays as to the extent that they conveyed important causal relationships necessary to address the essay prompt, and the system was trained to detect the ways that these were conveyed in natural language. While this approach is viable, a large corpus of training texts is required, and certainly more research is needed to identify the optimal strategies for determining if an essay conveys important concepts and relationships between them, as delineated by a documents model.

Magliano and Graesser (2012) argued that systems should ideally be developed that use multiple approaches for evaluating the relationships between student-generated content and semantic benchmarks, and the results of Wiley et al. (2017) support this. We have outlined several promising approaches in this chapter, but more research needs to be conducted to learn how to optimally combine the different approaches. One method would be to “let the data decide,” using a type of machine learning ensemble method called *stacking* (Wolpert, 1992), where the system learns how to balance different classification methods for different essays.

A third issue involves the potential to develop systems that can generalize to new prompts and document sets. This is indeed one goal of using machine learning to train systems to evaluate essays. However, given the idiosyncratic nature of the mapping between prompts, documents, documents models, and products for any given task, we see this as one of the more serious challenges raised in the last section. Any solution to this problem would require near-human level understanding of the texts. There are some advanced machine learning methods that use deep learning and unsupervised training (e.g., "zero-shot learning"; Norouzi et al., 2013), but these require extremely large corpora to have a chance at success, and are very much an area for future research.

Another potential future direction would be to be able to automatically assess the extent to which source reliability is considered when writing essays from sources that vary with regard to author credibility. Consider, for example, the 2016 election cycle in the United States and the preponderance of unreliable sources reporting "fake news." It is important for any consumer of information to be able to deal appropriately with varying levels of credibility. Especially in educational writing contexts, students are expected to explicitly identify sources. Future work could build on previous research which was effective at automatically identifying sourcing in texts (and the lack thereof) and also tutored students to improve their sourcing (Britt et al., 2004). Extensions to this research could focus on identifying statements that indicate the writer's evaluation of the sources.

Of course, this is challenging because of many factors. First, evaluating reliability requires that the information seeker has prior knowledge of what makes a good vs. poor source (akin to a schema for knowing where to find reliable information), and second, it would depend on a system that would be able to identify when/where unreliable information was present.

Another issue to consider is the extent that approaches have to be developed in a manner that is sensitive to the language. Systems that rely on keyword matching, regular expressions, and machine learning can be readily applied to just about any language system. However, systems that rely on high dimensional spaces, such as LSA and HAL require one to build the semantic spaces based on a large sample of documents (e.g., Landuaer & Dumais, 1997). As such, the semantic spaces that support these systems have to be built to support the linguistic contexts where the essay coding systems will be applied, but thankfully this is a viable endeavor (León, Olmos, Escudero, Cañas, & Salmerón, 2006; Olmos, León, Escudero, & Jorge-Botana, 2011). .

Finally, while our emphasis has been on the automatic evaluation of essays, we argue that the automatic assessment of other kinds of constructed responses, such as think aloud and question answering protocols (e.g., Magliano et al., 2011) has a utility in research on multiple document use. Higgs (2016) provided a proof of concept for this claim. Her study used LSA to compare think aloud protocols to semantic benchmarks derived from a documents model that enabled her to evaluate if readers tended to make intra and inter-text connections when reading documents. She found that readers tended to make connections within texts to information delineated as important in the documents model more so than connections across texts. Integration across texts likely happened after the initial readings of the documents in the set.

In summation, we hope that this chapter provides cognitive, learning, and educational scientists information about the tools and approaches they need to develop systems that have utility both in research and educational contexts. These systems will afford the use of essays and other student constructed responses to study multiple document use in the context of task-oriented reading (e.g., Rouet & Britt, 2011). However, as we have emphasized, they will also likely lead to the development of tools that could eventually be integrated into classroom use. Considerably more research is needed for that outcome to be realized and we hope that this chapter provides a foundation for this work.

References

Achieve (2013). Next generation science standards. Washington, DC: National Academies Press.

Aho, A. V. (1990). Algorithms for finding patterns in strings. In J. van Leeuwen (Ed.), *Handbook of theoretical computer science, volume A: Algorithms and complexity* (pp. 255–300). Location: Cambridge, MA: The MIT Press.

Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, *30*, 64-76.

Blaum, D., Griffin, T. D., Wiley, J., & Britt, M. A. (2017). Thinking about global warming: Effect of policy-related documents and prompts on learning about causes of climate change. *Discourse Processes, 54,* 303-316.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44,* 6–28.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*, 485–522.

Britt, M.A., & Rouet, J.-F. (2012). Learning with multiple documents: Component skills and their acquisition. In M.J. Lawson & J.R. Kirby (Eds.), *The quality of learning* (pp. 276-314). Cambridge UK: Cambridge University Press.

Britt, M. A., Wiemer-Hastings, P., Larson, A., & Perfetti, C. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14*, 359–374.

Chen, D., & Manning, C. D. (2014). *A fast and accurate dependency parser using neural networks*. Proceedings for Conference on Empirical Methods on Natural Language. Retrieved from http://emnlp2014.org/papers/emnlp2014-proceedings.pdf

Claassen, E. (2012). *Author representation in literary reading*. Utrecht, The Netherlands: John Benjamins.

Clark, K., & Manning, C. D. (2016, November). *Deep reinforcement learning for mention-ranking coreference models*. Paper presented at the Conference on Empirical Methods on Natural Language Processing, Austin Texas.

Council of Chief State School Officers. (2010). *The common core standards for English language arts and literacy in history/social studies and science and technical subjects*. Washington, DC: National Governors Association for Best Practices. Retrieved from http://www.corestandards.org

Dai, J., Raine, R.B., Roscoe, R., Cai, Z., & McNamara, D.S. (2011). The Writing-Pal tutoring system: Development and design. *Journal of Engineering and Computer Innovations*, *2*, 1-11.

de Marneffe, M. C., Connor, M., Silveira, N., Bowman, S. R., Dozat, T., & Manning, C. D. (2013, August). *More constructions, more genres: Extending Stanford dependencies*. Paper presented at the Interantional Conference on Dependency Linguistics, Prauge, Czech Republic.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis. Oxford: Philological Society: 1–32. Reprinted in F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952-1959. London: Longman.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111–127.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3,* 371-398.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, 36*, 180-193.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-95.

Hastings, P., Hughes, S., Magliano, J. P., Goldman, S. R., & Lawless, K. (2012). Assessing the use of multiple sources in student essays. *Behavior Research Methods*, *44*, 622-633.

Higgs, K. P. (2016). *Task specificity and multiple document integration* (Unpublished doctoral dissertation). Northern Illinois University, DeKalb, Illinois.

Hughes, S., Hastings, P., Britt, M. A., Wallace, P., & Blaum, D. (2015). Machine learning for holistic evaluation of scientific essays. In C. Conti, N. Hefferman, A. Mitrovic, & M. F. Verdejo (Eds), *Artificial Intelligence in Education,* (pp. 165-175). New York, NY: Springer.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55,* 534-552.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street ®: Computer-guided summary writing. In T. K. Landauer, D. M. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis* (pp. 263–277). Mahwah, NJ: Erlbaum.

Kopp, K., Rupp, K., Blaum, D., Wallace, P., Hastings, P., & Britt, M.A. (2016, November). *Assessing the influence of feedback during a multiple document writing task in science.* Poster presented at the Annual Meeting of the Psychonomic Society, Boston, MA.

Kurby, C.A., Britt, M.A., & Magliano, J.P. (2005). The role of top-down and bottom-up processes in between-text integration. *Reading Psychology*, *26*, 335-362.

Kurby, C.A., Wiemer-Hastings, K., Ganduri, N., Magliano, J.P., Millis, K.K., & McNamara, D.S. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods*, *35*, 244-250.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods*, *38*, 616-627.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.). *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, (pp. 660-665), Mahwah, NJ: Lawrence Erlbaum Associates.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, *44*, 608-621.

Magliano, J. P., McCrudden, M. T., Rouet, J. F., & Sabbatini, J. (in press). The modern reader: Should changes to how we read affect research and theory? In M. F. Schober, M. A. Britt, & D. N. Rapp (Eds.), *Handbook of discourse processes* (2nd ed.). New York: Routledge.

Magliano, J.P., & Millis, K.K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction*. *21*, 251-283.

Magliano, J. P., Millis, K.K., The RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the reading strategy assessment tool (RSAT). *Metacognition and Learning*, *6*, 131-154.

McCrudden, M. T., Magliano, J. P., & Schraw, G. (Eds.). (2011) *Text relevance and learning from text*. Greenwich, CT: Information Age.

McCrudden, M.T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review*, *19*, 113-139.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. In C. J. C. Burges, L Bottou, & M Welling (Eds) *Proceedings of Advances In Neural Information Processing Systems* (pp. 3111-3119). La Jolla, CA: Neural Information Processing Systems Foundation, Inc.

Millis, K.K., Magliano, J.P., Todaro, S., & McNamara, D.S. (2007). Assessing and improving comprehension with latent semantic analysis. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 207-225). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederikson, R. J. Mislevy, & I. I. Bejar (Eds.), *Tests theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Mitchell, T. (1997). *Machine Learning*. Columbus, OH: McGraw-Hill Education.

Morales, A., Premtoon, V., Avery, C., Felshin, S., & Katz, B. (2016). Learning to answer questions from Wikipedia Infoboxes**.** In *Proceedings of the 2016 Conference on Empirical Methods on Natural Language Processing (EMNLP 2016)*. Stroudsburg, PA: Association for Computational Linguistics.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., & Dean, J. (2013*). Zero-shot learning by convex combination of semantic embeddings*. Retrieved from https://arxiv.org/pdf/1312.5650.pdf

OECD (2008). PISA 2009 assessment framework - Key competencies in reading, mathematics and science. Paris: OECD. (Retrieved August 5, 2010 from http://www.oecd.org/).

Olmos, R., León, J. A., Escudero, I., & Jorge-Botana, G. (2011). Using latent semantic analysis to grade brief summaries: Some proposals. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*, 192-209.

Pellegrino, J. W., & Chudowsky, N. (2003). The foundations of assessment. *Interdisciplinary Research and Perspectives*, *1*, 103-148.

Pellegrino, J. W., & Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science of design of educational assessment*. Washington, DC: National Academy of Sciences.

Rouet, J.F. (2006). *The skills of document use: From text comprehension to web-based learning.* Mahwah, NJ: Erlbaum.

Rouet, J. F., & Britt, M. A. (2011). Relevance processing in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19-52). Greenwich, CT: Information Age.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Pearson Educational, Inc.

Shermis, M. D., & Burstein, J. (Eds.) (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York: Routledge.

Snow, C., & the RAND Reading Study Group. (2002). *Reading for understanding: Towards a R&D program for reading comprehension*. Santa Monica, CA: RAND.

Strømsø, H. I., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction*, *20*, 192-204.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Ventura, M. J., Franchescetti, D. R., Pennumatsa, P., Graesser, A. C., Hu, G. J., & Cai, Z. (2004). Combining computational models of short essay grading for conceptual physics problems. In J. C. Lester, R. M. Vicari, & F. Paraguac (Eds.), *Proceedings of the Intelligent Tutoring Systems Conference* (pp. 423–431). Berlin: Springer.

Vidal-Abarca, E., Mañá, A., & Gil, L. (2010). Individual differences for self-regulating task-oriented reading activities. *Journal of Educational Psychology*, *102*, 817-826.

Wiley, J., Goldman, S., Graesser, A., Sanchez. C., Ash, I., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal, 46,* 1060-1106.

Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., & Britt, M. A. (2017). Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education*, *23*, 1-33.

Wiley, J., & Voss, J.F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301-311.

Wineburg, S. S. (1991). On the reading of historical texts: Notes on the breach between school and the academy. *American Educational Research Journal*, *28*, 495-519.

Wolpert, D. H. (1992), "Stacked generalization". *Neural Networks, 5*, 241–259.

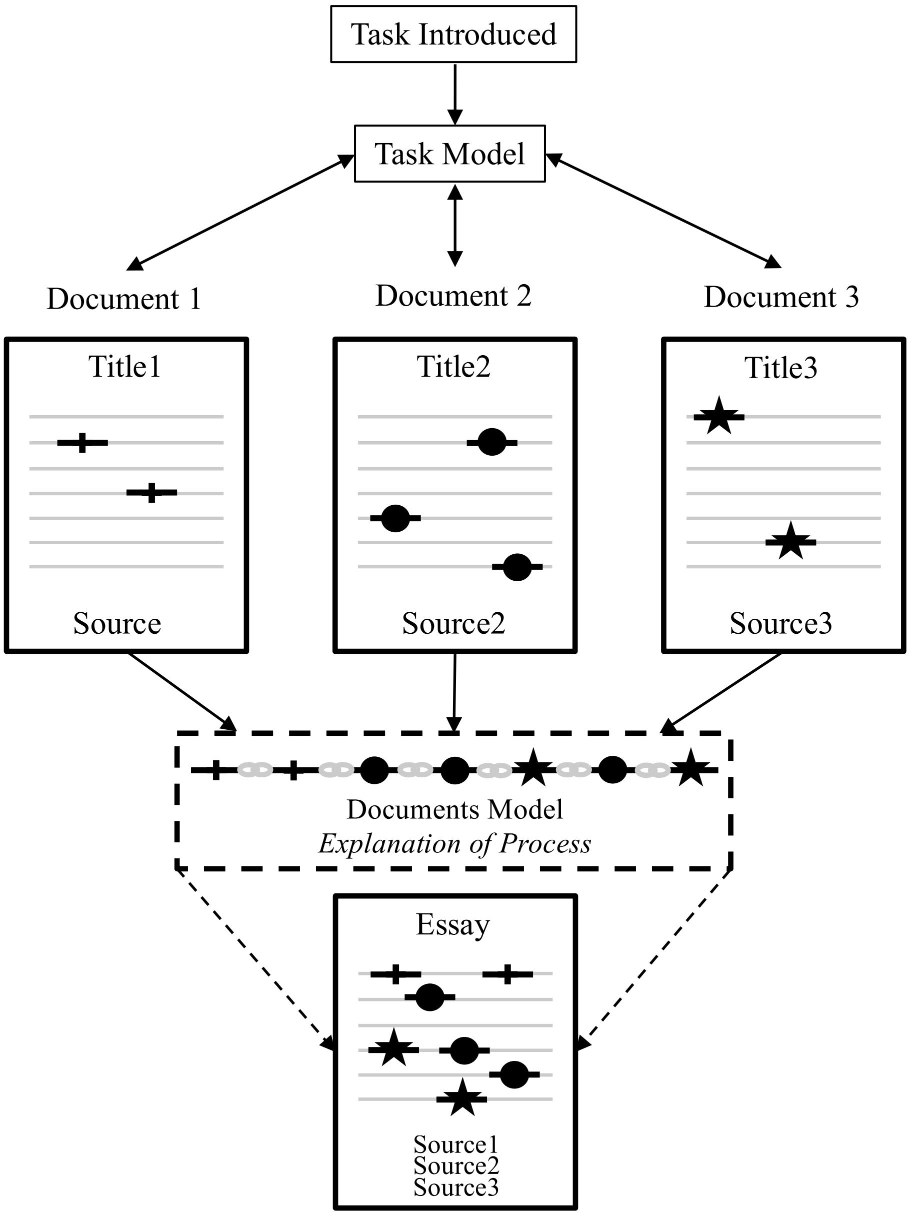
Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*, 292–297.

Table 1.

Example text idea units, participant essay idea units, and LSA cosines between the two.

Original Idea From Document	Protocols	LSA Cosines
"…ocean water temperatures increase by 3 to 5 degrees Fahrenheit"	Protocol 1a: "When the water temperature increases…"	0.83
	Protocol 2a: "The higher the water temperatures…"	0.83
	Protocol 3a: "The water temperature was around two to three degrees higher than normal…"	0.87
	Protocol 4a: "…allows for the water to become very warm."	0.77
	Protocol 5a: "…a large spike in heat …"	0.48
"…upsets the balance necessary for coral health."	Protocol 1b: "…upsets the balance for a healthy coral reef."	0.77
	Protocol 2b: "…declines the coral's overall health."	0.85
	Protocol 3b: "This ultimately deteriorates the corals overall health."	0.85
	Protocol 4b: "…providing large risks to the health and lives of the corals."	0.81
	Protocol 5b: "…brings the imbalance in the corals from keeping them thriving."	0.39

Figure 1. A graphic depiction of the nature of a multiple documents task.


Task Introduced
Task Model
Document 1
Document 2
Document 3
Title1
Source
Title2
Source2
Title3
Source3
Documents Model
Explanation of Process
Essay
Source1
Source2
Source3