DEEP AND SHALLOW NATURAL LANGUAGE UNDERSTANDING FOR IDENTIFYING EXPLANATION STRUCTURE

Peter Hastings, M. Anne Britt, Katheryn Rupp, Kristopher Kopp, & Simon Hughes

Introduction

Students' written explanations can shed light on their understanding of complex phenomena. If we can provide computational mechanisms to automatically analyze their explanations, we can better understand their thought processes, and help improve their understanding, either by providing feedback to teachers or through intelligent tutoring systems.

In this chapter, we describe deep and shallow processing of language by computers, but our research focus is on helping students learn how to read with deep comprehension. We define our view of deep comprehension in the context of argumentation and explanation from scientific texts, and we describe the supporting theoretical constructs. Then we present a study which assesses deep comprehension in student arguments. We demonstrate that shallow (computational) natural language understanding is sufficient to assess some aspects of student comprehension, but assessing deeper comprehension must be addressed by deeper computational approaches.

Deep and shallow Natural Language Understanding

Artificial Intelligence researchers use the term, "Natural Language Processing" (NLP), to refer to computational processing of human languages, as opposed to the processing of formal languages like computer programming languages. NLP has two major subfields: natural language understanding (NLU) and natural language generation (NLG). NLG is the production of texts or utterances in a human language to express intended meanings. NLU is the computational transformation from a human language into a representation which can be more easily manipulated ("understood") by computers allowing it to perform other tasks, for example, summarizing, answering questions, or simply adding to the system's "knowledge base".

In this chapter, we are focusing on Natural Language Understanding. Traditionally, this starts with parsing of texts into some formal representation. For example, with the sentence, "The dog ran after the bouncing toy," the part of speech for each word would be identified (article, noun, verb, etc.). Then, by applying grammar rules, the structure of the sentence could be derived (e.g., a noun phrase consisting of an article followed by a noun, and a verb phrase consisting of a verb and a noun phrase). Then that structure could be converted to a logical statement of the sentence's meaning (CHASE (DOG, BOUNCING-BALL)), and pragmatic processing would be applied to derive the contextualized meaning (e.g., What were the intentions of the speaker? Which dog is this referring to?).

Unfortunately, the ambiguities in human language make this process very computationally expensive. For example, "dog" and "toy" could both be nouns or verbs, and "bouncing" could be used as a verb, noun, or adjective. The ambiguities multiply when different grammatical rules can apply (e.g., prepositional phrases which can attach to multiple antecedents). These ambiguities could be resolved by applying contextual and world knowledge — humans do it all the time. But computers struggle with it, because it is practically impossible to predict all such knowledge that might be applicable. This type of deep NLU provides a representation of the language which is flexible, but *brittle* — unexpected words or syntax cause complete failure.

Shallower NLU approaches provide broader coverage of texts, but at the cost of loss of information. For example, keyword-based methods are based on the presence of particular words or sets of words. Many such techniques ignore word order (and, therefore, the syntactic structure) of the text. Some focus primarily on the descriptive statistics of the texts (e.g., counts of different types of words). There is a wide range of depth in NLU techniques, and, although that depth is not easily quantifiable, the tradeoff between coverage and depth is common.

Deep and shallow language processing in educational contexts

Educational contexts provide the motivation for our research and for many others', including most of the other research in this volume. It also allows us to explore the relationship between the depth of understanding of a "teacher" (which could be a human or a computational tutor) and that of a student. Does a teacher need to deeply understand a student's text in order to help the student learn more deeply?

Studies of untrained human tutors showed that they did not deeply analyze what their students said, but still helped the students learn much more effectively than classroom instruction because they used a variety of prompting strategies that kept the discussion going (Graesser, Person, & Magliano, 1995). This provided the inspiration for the AutoTutor family of dialog-based intelligent tutoring systems, which use rather shallow NLU, but have been shown to be very successful at helping students learn in both lab settings (Nye, Graesser, & Hu, 2014) and as an adjunct to classroom teaching (Arnott, Hastings, & Allbritton, 2008).

Is there a limit to the depth of understanding that this shallow NLU approach can foster? We explore that question in this chapter and present a study which sheds light on the answer.

Deep Comprehension from science texts

Distinguishing "deep" from "shallow" comprehension

In the past 15 years, there has been a growing call to expand the definition of reading beyond learning to decode and comprehend an author's meaning (Britt, Rouet, & Durik, 2017; Graesser, Chipman, Leeming, & Biedenbach, 2009; Institute of Education Sciences, 2010; OECD, 2015; Snow and the Rand Reading Group, 2002). Graesser et al. (2009) state that deep comprehension involves "an analysis of causal mechanisms, logical

explanations, creation and defense of arguments, management of limited resources, tradeoffs of processes in a complex system, and a way to resolve conflicts" (p. 84). This is in contrast to shallow learning which involves "perceptual learning, motor skills, definitions of words, properties of objects, and memorization of facts" (p. 84).

In this chapter, we define deep comprehension as going beyond memory for explicitly presented facts to being able to reconstruct explanations for scientific phenomena and representing the evidence for one or more, possibly competing, explanations. In contrast, we view shallow comprehension as learning and memorizing new vocabulary and facts from texts (Graesser et al., 2009).

Structure of scientific explanations

Scientific explanation and argumentation have been identified in the Next Generation Science Standards (NGSS) as foundational skills that enable students to think critically about natural phenomena and engage in authentic practices of science (Achieve, 2013). Students need to understand explanations of phenomena in the natural world, to evaluate which explanations are best supported, and to communicate those ideas to other people (Achieve, 2013; Britt, Richter, & Rouet, 2014; OECD, 2015; Osborne & Patterson, 2011). Most students need training in explanation and argumentation in science (Duschl & Osborne, 2002; Hastings, Hughes, Blaum, Wallace, & Britt, 2016). The focus of this chapter is limited to students' understanding of explanations from written texts.

While texts or segments of texts can include several types of structures, including description, collection, sequence and causation (Meyer & Freedle, 1984), the two that are the hallmark of deep comprehension in science texts are causal explanation and argumentation. While there are several types of explanations (Braaten, & Windschitl, 2011; Hempel & Oppenheim, 1948; Salmon, 1989), the U.S. science standards focus on students developing skill in representing and using causal explanations (Achieve, 2013). This includes reading to understand "how" and "why" questions about phenomena such as "How do coral get bleached?" and "Why do humans get skin cancer?". Causal explanations for scientific phenomena have a general structure that includes one or more initiating causes coherently connected to intervening states or events that lead to the to-be-explained outcome (Achieve, 2013, Chinn & Malhotra, 2002; Hempel & Oppenheim, 1948). For instance, an explanation for "what causes coral bleaching" could include "that changes in water temperature in the ocean caused by changing winds can lead to less carbon dioxide in water which is necessary for photosynthesis so that the algae will survive. Without these algae, the coral will die but before they die, they become white (a process called coral bleaching)". Here the initiating cause is changes in wind patterns that changes water temperature and the to-be-explained-outcome is coral bleaching. An even more complete explanation would include a second initiating cause of storms leading to changes in the salinity of the water.

Assessing deep structure of explanations

First, we briefly describe a new Evidence-based Argument assessment of deep comprehension of scientific explanations that has students read a set of documents to answer a "how" or "why" question about the causes of a scientific phenomenon by

integrating across multiple sources (Goldman, et al., submitted; Hastings et al., 2016; Wiley et al., 2009). We present this new assessment in some detail because it is the only one we know of that examines explanation understanding from multiple documents. Then we present results from a study using this assessment to show the importance of identifying explanation structure as an indicator of the student's deep comprehension in addition to identifying content (more shallow comprehension).

In this assessment, a document set was built around a causal model, shown in Fig. 1, for the coral bleaching topic. To answer the question, the reader had to create inferences based on their knowledge of what an explanation is. To construct a complete, integrated, coherent explanation, a reader would have to identify all potential initiating causes and then make explicit all the relationships between concepts that are linked to them, paying special attention to intermediate processes or mechanisms. In contrast to a shallower task which would present a single text that has a complete and coherent explanation, students had to integrate information from multiple documents. Our assessment of deep comprehension included both representing (as measured by multiple choice items) and communicating the structure of an explanation (as measured by an essay).

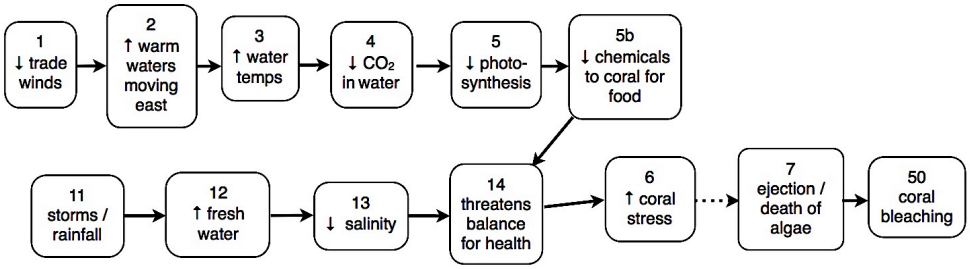

Figure 1: Causal model for coral bleaching

In a recent study, college students were given the coral bleaching assessment (Kopp et al., 2016). One document gave general background information, three documents described different parts of the explanation, and one was a graph of the relationship between one initiating factor (trade winds) and coral bleaching over the last 30 years. The students were asked to read the documents and then write an essay (with the documents present) that answers the question "What leads to differences in the rates of coral bleaching?"

Before reading, all participants were told what we meant by a causal explanation in science, including information about coherence ("a detailed series of important factors", "illustrates the intermediate steps that lead to") and completeness ("there are multiple causes", "so please be as complete as possible"). These instructions were also intended to help them understand that they needed to integrate across documents. To make sure they encoded the instructions, we presented the instructions with key phrases as a closed task (i.e., fill in blanks).

Participants then typed their essays while reading paper copies of the documents. After finishing their first draft, the experimenter put it into the "computer annotator" which

automatically scored it and provided feedback. In actuality, their feedback was randomly determined but they were led to believe that the computer scored it. They were given one of four types of feedback. The two conditions that we are focused on here are the Revise feedback and the Complete feedback conditions. All feedback told them that this was a "good start" and that they would now be given a chance to go back and revise their essay. For the Revise feedback condition, this was all they were told. For the Complete feedback condition, they were reminded about coherence (i.e., "You should include factors from different sources and explain how they interact and cause coral bleaching") and completeness (i.e., "there may be multiple causal chains. You may want to now focus on how different factors from different sources interact and create a separate causal chain").

In addition to the essay task that provided a measure of students' ability to communicate the structure of an explanation (participant's draft) and revise this communication based on feedback (participant's revision), we also had several measures of learning that were less production heavy. These included 9 multiple-choice items to test their understanding of specific parts of the causal explanation from within a single document and for links across documents. For additional measures of deep comprehension (i.e., critiquing and evaluating models) in this novel assessment see Goldman (under review). In the next section, we describe how these essays were scored by humans and use this study to illustrate two conditions in which scoring for structure could enhance a real computer assessment tool.

Description of human scoring

The initial essays were annotated by human scorers according to the causal model. As shown in Fig. 1, there were 13 causal elements (e.g., higher water temperature) that could be connected and two distinct initiating factors (i.e., reversing trade winds, storms). As in our prior work (Hastings, et al., 2016; Wiley et al., 2017), we scored the essays for the number of Unique concepts mentioned from the causal model regardless of how they were connected. This measures important information selection but not how that information was structured. To get at explicit structural relations to form a coherent, well-connected explanation, we scored the essays for the number of Intervening factors they contained. That is, the number of concepts that the student explicitly mentioned as connected from an initiating cause to the outcome. Both of these methods were useful in scoring explanation essays. In one study, we found that both measures predicted middle school and high school students' learning of the causes of global warming from multiple documents as measured by an inference verification task (Wiley et al., 2017).

Results showing importance of structure scoring

We first looked at how performance on the initial drafts would predict learning, collapsing across conditions because it was not significant and the conditions were identical at this point. The number of Unique concepts and Intervening factors both significantly correlated with multiple-choice performance, $r = .32$, $p < .05$, $r = .47$, $p < .001$, respectively. Multiple regression was used to examine the extent to which Intervening factors (structure) predicted multiple-choice performance (learning) beyond the Unique concepts (content selection).

In the first step of the regression, Unique concepts were a significant predictor of learning, accounting for 10% of the variance ($F(1, 43) = 4.86$, $p < .05$ for change in R). Most importantly, adding Intervening factors on the second step significantly increased the amount of variance accounted for to 22% ($F(1, 42) = 6.44$, $p < .05$ for change in R). Thus, at least for college students that are told what an explanation is, coding essays for structure does help to identify those who learned more on their initial drafts beyond just coding for concepts.

Next, we consider whether identifying explanation structure enables the creation of feedback which is beneficial for students' revisions of their essays. For this analysis, we assessed the number of Unique concepts and Intervening factors for the revised explanations (i.e., after they received feedback) and looked at performance on the revisions between conditions. An ANCOVA with unique concepts from the initial draft as the covariate showed that there was a significant difference in conditions for the number of unique concepts on the revision. The Complete feedback condition produced significantly more Unique concepts than the Revise condition: 6.8 (SE = .30) > 5.7 (SE = .33); $F(1, 42) = 5.80$, $p = .02$, (MSE = 2.14), $d = .77$. This shows that students selected additional content to include in their essays which would suggest the feedback improved learning of the explanation. The two conditions, however, did not statistically differ for the number of Intervening factors in their revised essays: Complete = 2.7 (SE = .30) vs. Revise = 2.3 (SE = .34); $F(1,42) = .85$, $p > .05$, (MSE = 2.26), d=.48. This shows that the feedback did not lead to better structured essays. Therefore, if only the amount of content were considered, the conclusion would be that the structure feedback was effective. However, if the goal is to help students integrate information into a coherent explanation for a phenomenon, the conclusion would be that the feedback was *not* effective. Only by performing a deep analysis of explanation structure is this failure evident.

Computational approaches for assessing structure

The main goal of this chapter is to address the question of which computational techniques (NLU) are effective for determining the causal structure in student explanations. First, we need to distinguish this task from the more typical task of applying a single holistic score which evaluates the explanation quality. A holistic score could be based on a number of factors which are aggregated over the explanation, for example, the number of important terms from the sources which the students read, and the number of "causal" terms like "causes", "leads (to)", and "then". Instead, causal structure identification requires locating specific concepts from the causal model in the student's explanation, and determining how they are explicitly connected to the target outcome. From this, we can create a model of the explanation that is a subgraph of the complete model. This can provide the basis for very specific feedback on the student's explanation, including the content that was included, that which was omitted, and how they were connected. Prior research shows that making suggestions based on specifics of student texts results in significant improvements in the quality of revisions (Britt, Wiemer-Hastings, Larson, & Perfetti, 2004).

In this section, we describe the computational approaches which have been shown to be effective at identifying the specific structure of student explanations, including

identification of both concepts and causal connections. We also describe some techniques that are not well-suited for this task. For the effective approaches, we provide information about specific "off-the-shelf" tools that can give the required computational analyses, and also describe some utilities for creating data that are needed for these tools.

Shallow NLU for structure identification

As previously mentioned, some shallow NLU techniques are suitable for holistic scoring of explanation quality, but they are limited in their ability to identify causal structure. For example, keywords for a particular topic could be identified, and very simple search mechanisms could count the occurrence of those keywords in the students' explanations. Causation-related keywords can also be searched in the explanations. A holistic score for each explanation can be assigned based on the ratio of these keywords to the total number of words, perhaps also using the total number of words or sentences in the explanation or similar factors as additional indicators. But such an approach cannot identify the specific causal structure embedded within the explanations.

The Coh-Metrix system (available at http://cohmetrix.com) goes significantly beyond keyword search. It provides 108 different measures of a submitted text, including descriptive statistics (number of words, sentences, paragraphs, etc.) and measures related to readability, cohesion and text complexity (McNamara & Graesser, 2012). Some of these measures are derived from special-purpose collections of word classes developed for the system. Others come from more general-purpose tools. Although Coh-Metrix does provide measures that include more local information, for example, the average number of sentences in the text with nouns in common between two adjacent sentences, the measures are all aggregates over the entire text. Thus, they can be effective at providing a holistic score for an explanation (Wiley et al., 2017), but they are inappropriate for use in identifying the structure.

Used appropriately, however, some shallow NLU techniques can identify some aspects of explanation structure. For example, a keyword match or regular expression match (i.e., matching patterns of word combinations with alternatives) can be applied to successfully identify concepts from the source documents in student explanations. Using a tool to aid creation and testing of such patterns, Hastings et al. (2012) showed success in identifying very specific concepts from the texts, but were less successful with higher-level concepts, and concepts that required integration across documents. This type of technique was also incapable of identifying causal claims between the concepts due to the wide variety of ways that the students expressed those connections.

Word semantic methods

The keyword matching method described above relies on identifying specific words in the student text. It is a shallow method; if the exact word does not appear (due, for example, to misspelling or use of a synonym), the corresponding concept is not identified. Regular expression matching is slightly less brittle because it allows more alternatives and combinations. But both are dependent on the exact words as they are included in the text. If the system had a deeper representation of word meanings, it could handle more variation in terms of word synonyms and syntactic constructions. Word semantic methods can provide this deeper representation.

Two well-known examples of such methods are Latent Semantic Analysis (LSA, Landauer & Dumais, 1997), and word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Both systems represent words as vectors in a high-dimensional space, typically using 50 to 300 dimensions. Both create their representations from a corpus of texts. LSA uses a mathematical reduction of dimensions from a word-by-document co-occurrence matrix. Word2vec is a neural network technique that is trained to predict the probability of a word from the context of the words that surround it. The key advantage of these types of methods is that their representation places words with similar meanings nearby each other within the high-dimensional space.

LSA, available at http://lsa.colorado.edu, has been used in many ways, for example, identifying synonyms, performing analogies, and selecting related texts. In the context of causal structure identification, the most obvious way to use it is to compare the similarity of the student's explanation to the original source documents. The trick is to find the right level of granularity to compare them. With LSA, word vectors can simply be combined to create vectors for phrases, sentences, etc. The geometric cosine between vectors gives a measurement of how close two vectors are, ranging, in practice, from 0 to 1. LSA is a "bag-of-words" model; it completely ignores word order and syntax. So how does one know which sets of words to compare? If each concept from the causal model is identified by a single word, then each of those words can be compared with every word in the student's explanation to locate the concepts. But then, its performance is very similar to keyword matching. At the other extreme, if the whole student explanation is compared to the entire text of the source documents, the cosine score can be used as a general holistic score for the similarity, but it is bound to be a rough estimate at best due to extra content in the source documents that is not part of the causal model. A common intermediate technique is to do sentence-by-sentence matching (Hastings et al., 2012). However, because the concepts and causal connections are normally represented by smaller phrases, this technique can only accurately identify single-sentence causal connections in the student explanations which were given as single-sentence connections in the source documents.

Code for training word2vec is available on the web in many languages, including C, Python, and Java. Word2vec can be used in the ways described above as well, but it is more often used as a rich input to other neural network techniques. More about these below.

Machine learning approaches

When working in an educational context, each topic and task may have a different sub-language associated with it. In other words, student causal explanations for one scientific domain will include different words and possibly even different syntactic constructions than those in another domain. And other genres of texts will be expressed in different ways even if they are on the same topic (Goldman et al., 2016). Because of this variability, the best way to ensure deep analysis of student texts is to use machine learning approaches to customize the structure analysis to the particular task and topic. That can, however, require more sophisticated techniques. But here also there is a range of depth of mechanisms available, some easier to implement than others. In this section, we describe some techniques that have been shown to be effective for identifying

structure in causal explanations.

Text annotation for machine learning

The most directly applicable machine learning methods for structure identification are supervised learning methods, that is, they are trained from a set of texts in which the target information has been previously marked by expert coders. Here also, the granularity of analysis is important. A simple, easily accessible technique for annotation is to put each student explanation into a large spreadsheet, with one sentence in every row, and, in separate columns, identify the concepts that occur in that sentence and the causal connections if any. The utility of such an approach is very limited, however. It could be used to evaluate a sentence-based LSA approach, but, because the specific words within the sentence that identify the concepts and causal connections are not identified, it cannot be used to train a deeper machine learning approach.

Instead, an annotation tool like brat (http://brat.nlplab.org/, Stenetorp et al., 2012) is recommended. This type of tool makes it easy to label the specific words that correspond to important concepts and causal connections in the texts and to visualize the contents of each text, and it stores the annotations in a simple, text-based representation. Fig. 2 shows an example of annotation in brat from a biomedical domain.

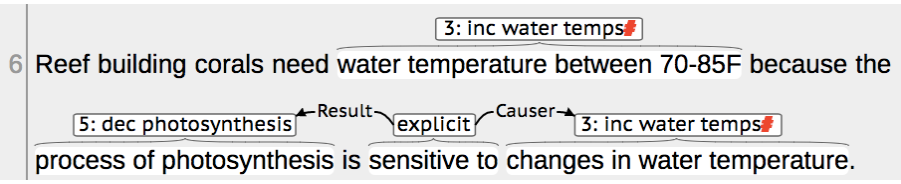

Figure 2: An example of annotation with brat

Multi-word

A very simple way to use machine learning for causal structure identification is to automatically learn the type of regular expressions described above. When learning new patterns, there is a trade-off between coverage of the patterns (the hit rate) and the number of false alarms. If you make the pattern more general, you get more hits, but the number of false alarms increases as well. Hastings, et al. (2012) took an iterative approach to creating patterns, adding different options for concepts until the combined accuracy went down. They found that, while the technique was not best overall, compared to one based on LSA and one based on another machine learning method (SVMs, described below), this approach was superior for identifying causal connections that students made between source documents, as opposed to those where the cause and the effect were both described within the same source document. This type of causal connection was relatively rare in the student explanations, and that may be one reason for the advantage of the multi-word approach here. The other learning methods need more examples to successfully distinguish texts which do and do not contain these connections.

Support Vector Machines (SVMs)

SVMs (Vapnik & Chervonenkis, 1971; downloadable code at http://svmlight.joachims.org) provide a considerably more sophisticated machine learning

technique and are very popular for many applications, including text classification. Explanation structure identification is treated as a classification problem by posing the question, "Does a particular sentence (or phrase, or paragraph, or some other amount of text) include a concept or causal connection from the causal model?" For this type of task, typically an SVM will be trained with each word in the training corpus treated as a separate input feature or with ngrams as features. During the training, the SVM learns to separate the positive from the negative examples, i.e., the examples containing the target concept or causal connections from those that do not. For many tasks, SVMs perform very well. But their depth of analysis for causal structure identification is limited in two ways. First, the technique is a bag-of-words model, ignoring word order and syntactic structure. Second, as with the LSA approach described above, it is not clear how much text (word, sentence, etc.) to compare. For example, if a student sentence is compared to a source sentence and no matching causal connection is found, that could be because there is none there to be found, or it could be because there are additional words in the sentence which outweigh the contribution of the causal connection words. Nevertheless, Hastings et al. (2012) showed that an SVM provided better overall classification than the multi-word approach and an LSA sentence-to-sentence comparison approach.

Window-based tagging

The techniques described so far have either searched for specific keywords or patterns in student explanations, or they have used deeper representations to compare relatively arbitrary spans of text (usually sentences). One way to combine more localized search with a more robust, learned representation is called window-based tagging (Hughes, Hastings, Britt, Wallace, & Blaum, 2015). In this technique a window of words with a fixed, odd size is "slid" across the text to create training examples where the goal is to predict the classification (concept or causal connector) of the central word in the window. The features are the other words in the window along with their relative distance before or after the target. As an example, consider a window of size 5, and an example sentence, " . . . which causes the salinity to drop, this drop . . . ", where the phrase, "salinity to drop" is marked with code #13 in the graph above. The training instances from this example would include: [which:-2, causes:-1, salinity:+1, to:+2] → none; [causes:-2, the:-1, to:+1, drop:+2] → 13; etc., where the items in brackets are the 4 features including relative position, and the item after the arrow is the predicted classification. Using these features as input to a set of SVM classifiers[1], one for each concept, Hughes et al. (2015) achieved a high level of accuracy, $F_1 = 0.85$[2], at identifying concepts in sentences. The accuracy for causal connections was also quite good, $F_1 = 0.65$, but was limited because it depends on the results of the concept classification.

---

[1] For this evaluation, the bigrams were also included, i.e., two consecutive words along with their relative positions.

[2] F_1 ranges from 0 to 1, and is defined as *(2*Precision*Recall)/(Precision + Recall)*, where *Recall = Hits / (Hits + Misses)*, and *Precision = Hits / (Hits + False Alarms)*.

Neural networks and deep learning

Advances in algorithms and computational power have recently enabled deep neural networks to accomplish a wide variety of tasks including natural language processing tasks. At a high level, the idea is this: a network of very simple units takes inputs (either in groups, or one at a time) and, via weighted connections to other units, gradually learns to improve performance on some task by adjusting the weights to reduce overall error. One of the most successful approaches for this type of problem is Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (Goodfellow, Bengio, & Courville, 2016). It can "read" a text one word at a time, adding information about the word to its current state. It can also "decide" when to output the state information and when to clear it out. All of the decisions and the way new information is added are controlled by weights which are learned during the training process.

Using syntactic structure

As mentioned at the beginning of the chapter, the traditional method of natural language understanding starts with determining the syntactic structure of the text. Intuitively, this makes sense: the syntax is the framework that the meaning is built on. However, the ambiguities in language make all but the simplest forms computationally expensive to derive. And circularities often occur: you can't determine the meaning without the syntax, but you can't determine the syntax without the meaning. Nevertheless, there are fast statistical parsers available like the Stanford Parser (Socher, Bauer, Manning, & Ng, 2013, https://nlp.stanford.edu/software/lex-parser.shtml) or SyntaxNet (https://research.googleblog.com/2016/ 05/announcing-syntaxnet-worlds-most.html) which are trained from large corpora of texts, that can do a good job of determining sentence structure. The performance of these systems depends on the size of the training set, so the default choice is to use them with a pre-trained corpus, typically newspaper texts. If the language that the students uses differs significantly from these texts, which it often does, the parser may be less successful.

Hybrid approaches

Hastings et al. (2012) compared several different approaches for analyzing causal explanation structure, pattern matching, an LSA-based approach, and a sentence-level machine learning approach. Their results supported the "functional semantic overlap hypothesis", namely, that each different type of method analyzed had strengths and weaknesses for identifying different types of concepts. Thus, a hybrid approach could be very effective. The trick to making a hybrid approach work is to know when to prefer the classification of one method over the other. An ensemble learning (or multiple classifier) method (Rokach, 2010) is trained to make just this type of determination. There are several variants on this approach, but with all of them, the trained system learns which of several different methods to "trust" in different situations.

Conclusions

How deeply do students need to comprehend texts? That depends largely on the task for which they are reading (Britt et al., 2017). Likewise, the types of computational approaches appropriate for analyzing student texts depends on the goals of that analysis. Shallow NLU approaches have been shown to be very effective at analyzing texts in a

variety of settings that promote student learning. To effectively uncover deeper aspects of texts like the causal structure in student explanations of scientific phenomena, more advanced NLU techniques are required.

In this chapter, we have shown why it is important to teach students to be able to read with deep comprehension, and why deep NLU is important for helping students learn to do that. And we have described a variety of NLU approaches, both shallow and deep, that can support student learning.

References

Achieve, Inc. (2013). *Next Generation Science Standards*. Washington, D.C.: National Academies Press.

Arnott, E., Hastings, P., & Allbritton, D. (2008). Research methods tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods, 40*(3), 694–672.

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education, 95*, 639-669.

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.

Britt, M.A., Richter, T., & Rouet, J.-F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104-122.

Britt, M.A., Rouet, J.-F., & Durik, A.M. (2017). *Literacy beyond text comprehension: A theory of purposeful reading*. New York: NY: Routledge.

Chinn, C.A., & Malhotra, B.A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175-218.

Duschl, R., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*, 39-72.

Goldman, S.R., Britt, M.A., Brown, W., Cribb, G., George, M., Greenleaf, C., Lee, C.D., Shanahan, C. & Project READI. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 51*, 1-28.

Goldman, S.R., Greenleaf, C., Yukhymenko-Lescroart, M., Brown, W., Ko, M., Emig, J., George, M.A., Wallace, P., Blaum, D., Britt, M.A., and Project READI. (submitted). Explanatory modeling in science through text-based investigation: Testing the efficacy of the project readi intervention approach.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. (http://www .deeplearningbook.org)

Graesser, A.C., Person, N.K., & Magliano, J.P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 359-387.

Graesser, A.C., Chipman, P., Leeming, F., & Biedenbach, S. (2009). Deep learning and emotion in serious games. In U. Ritterfield, M. Cody & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects* (pp. 81-100). Mahwah, NJ: Routledge.

Hastings, P., Hughes, S., Britt, M.A., Wallace, P., & Blaum, D. (2016). Stratified learning for reducing training set size. In Proceedings of the 13th *International Conference on Intelligent Tutoring Systems, ITS 2016* (p.341 - 346). Berlin: Springer.

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

Hempel, C.G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, XV*, 135–175.

Hughes, S., Hastings, P., Britt, M.A., Wallace, P., & Blaum, D. (2015). Machine learning for holistic evaluation of scientific essays. In Proceedings of *Artificial Intelligence in Education 2015*. Berlin: Springer.

Institute of Education Sciences. (2010). *Request for applications: Reading for Understanding research initiative*. CFDA Number: 84.305F.

Kopp, K., Rupp, K., Blaum, D., Wallace, P., Hastings, P., & Britt, M.A. (Nov, 2016). *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.

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

McNamara, D.S., & Graesser, A.C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P.M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: IGI Global.

Meyer, B.J.F., & Freedle, R.O. (1984). Effect of discourse type on recall. *American Educational Research Journal, 21*, 121-143.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Nye, B.D., Graesser, A.C., & Hu, X. (2014). AutoTutor and family: A review of 17 years

of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*, 427-469.

OECD (2015) *PISA 2015 draft frameworks*. http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm

Osborne, J.F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education, 95*, 627-638.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*, 1–39.

Salmon, W.C. (1989). Four decades of scientific explanation. *Scientific Explanation*, *13*, 3-219.

Socher, R., Bauer, J., Manning, C.D., & Ng, A.Y. (2013). Parsing with compositional vector grammars. In Proceedings of the 51st Annual Meeting of the *Association for Computational Linguistics,* 1, (p. 455-465). Sofia, Bulgaria.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012, April). brat: a web-based tool for NLP-assisted text annotation. In Proceedings of the *Association for Computational Linguistics demonstrations session at EACL 2012*. Avignon, France. Retrieved from http://brat.nlplab.org

Vapnik, V., & Chervonenkis, A., (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and Its Applications*, Volume XVI, Number 2, pages 264–280, 1971.

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., Hughes, S., Wallace, P., Griffin, T.G., & 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*, 27, 758–790.