Identifying science concepts and student misconceptions in an interactive essay writing tutor

Steven Bethard
University of Colorado
Boulder, Colorado, USA
steven.bethard@colorado.edu

Ifeyinwa Okoye
University of Colorado
Boulder, Colorado, USA
ifeyinwa.okoye@colorado.edu

Md. Arafat Sultan
University of Colorado
Boulder, Colorado, USA
arafat.sultan@colorado.edu

Haojie Hang
University of Colorado
Boulder, Colorado, USA
haojie.hang@colorado.edu

James H. Martin
University of Colorado
Boulder, Colorado, USA
james.martin@colorado.edu

Tamara Sumner
University of Colorado
Boulder, Colorado, USA
tamara.sumner@colorado.edu

Abstract

We present initial steps towards an interactive essay writing tutor that improves science knowledge by analyzing student essays for misconceptions and recommending science webpages that help correct those misconceptions. We describe the five components in this system: identifying core science concepts, determining appropriate pedagogical sequences for the science concepts, identifying student misconceptions in essays, aligning student misconceptions to science concepts, and recommending webpages to address misconceptions. We provide initial models and evaluations of the models for each component.

1 Introduction

Students come to class with a variety of misconceptions present in their science knowledge. For example, science assessments developed by the American Association for the Advancement of Science (AAAS)[1] showed that 49% of American 6th-8th graders believe that the Earth's tectonic plates are only feet thick (while in fact they are miles thick) and that 48% of American 6th-8th graders believe that atoms of a solid are not moving (while in fact all atoms are in constant motion). A key challenge for interactive tutoring systems is thus to identify and correct such student misconceptions.

In this article, we develop an interactive essay writing tutor that tries to address these challenges. The tutor first examines a set of science webpages to identify key concepts (Section 4) and attempts to order the science concepts in a pedagogically appropriate learning path (Section 5). Then the tutor examines a student essay and identifies misconception sentences (Section 6) and aligns these misconceptions to the true science concepts (Section 7). Finally, the tutor suggests science webpages that can help the student address each of the misconceptions (Section 8).

The key contributions of this work are:

- Demonstrating that a summarization approach can identify core science concepts
- Showing how a learning path model can be bootstrapped from webpages with grade metadata
- Developing models for misconception identification based on textual entailment techniques
- Presenting an information retrieval approach to aligning misconceptions to science concepts
- Designing a system that recommends webpages to address student misconceptions

2 Related work

Interactive tutoring systems have been designed for a variety of domains and applications. Dialog-based tutoring systems, such as Why2-Atlas (VanLehn et al., 2002), AutoTutor (Graesser et al., 2004) and MetaTutor (Azevedo et al., 2008), interact with students via questions and answers. Student knowledge is judged by comparing student responses to knowledge bases of domain concepts and misconceptions. These knowledge bases are typically manually curated, and a new knowledge base must be constructed for each new domain where the tutor is to be used.

[1] http://assessment.aaas.org/

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 12–21,
Montréal, Canada, June 3-8, 2012.

Essay-based tutoring systems, such as Summary Street (Wade-Stein and Kintsch, 2004) or CLICK (de la Chica et al., 2008b), interact with students who are writing a summary or essay. They compare what the student has written to domain knowledge in the form of textbooks or webpages. They typically do not require a knowledge base to be manually constructed, instead using natural language processing techniques to compare the student's essay to the information in the textbooks or webpages.

The current work is inspired by these essay-based tutoring systems, where interaction revolves around essay writing. However, where Summary Street relies primarily upon measuring how much of a textbook a student essay has "covered", we aim to give more detailed assessments that pinpoint specific student misconceptions. CLICK targets a similar goal to ours, but assumes that accurate knowledge maps can be generated for both the domain knowledge and for each student essay. Our approach does not require the automatic generation of knowledge maps, instead working directly with the sentences in the student essays and the webpages of science domain knowledge.

3 System overview

Our system is composed of five key components. First, a core concept identifier examines domain knowledge (webpages) and identifies key concepts (sentences) that describe the most important pieces of knowledge in the domain. Second, a concept sequencer assigns a pedagogically appropriate order in which a student should learn the identified core concepts. Third, a misconception identifier examines the student essay and identifies sentences that describe misconceptions the student has about the domain. Fourth, a misconception-concept aligner finds a core concept that can be used to correct each misconception. Finally, a recommender takes all the information about core concepts and student misconceptions, decides what order to address the misconceptions in, and identifies a set of resources (webpages) for the student to read.

To assemble this system, we draw on a variety of existing datasets (and some data collection of our own). For example, we use data from an annotation study of concept coreness to evaluate our model for identifying domain concepts, and we use data from science assessments of the American Association for the Advancement of Science to train and evaluate our model for identifying misconceptions. We use this disparate data to establish baseline models for each of the tutor's components. In the near future, this baseline tutoring system will be used to collect student essays and other data that will allow us to develop more sophisticated model for each component.

4 Identifying core concepts

This first module aims at automatically identifying a set of *core concepts* in a given set of digital library resources or webpages. Core concepts in a subject domain are critical ideas necessary to support deep science learning and transfer in that domain. From a digital learning perspective, availability of such concepts helps in providing pedagogical feedback to learners to support robust learning and also in prioritizing instructional intervention (e.g., deciding the order in which to treat student misconceptions). A concept can be materialized using different levels of linguistic expressions (e.g. phrases, sentences or paragraphs), but for this work, we focus only on individual sentences as expressions of concepts.

We used COGENT (de la Chica et al., 2008a), a multi-document summarization system to extract concepts (i.e. sentences) from a given set of resources. In the following two subsections, we describe the COGENT system, discuss how we used it for core concept extraction and report the results of its evaluation of effectiveness.

4.1 Model

COGENT is a text summarizer that builds on MEAD (Radev et al., 2004), a multidocument summarization and evaluation platform . MEAD was originally developed to summarize news articles. COGENT aims to generate pedagogically useful summaries from educational resources.

COGENT extends MEAD by incorporating new features in the summarization process. MEAD uses a set of generic (i.e. domain-independent) features to evaluate each sentence in the given set of documents. These features include the length of the sentence, the distance from the sentence to the beginning of the document, etc. Individual scores of a sentence along

these dimensions are combined to assign a total score to the sentence. After removing redundant sentences, MEAD then generates a summary using the sentences that had the highest scores. A user-specified parameter determines the number of sentences included in the summary.

COGENT extends this framework by incorporating new domain-general and domain-specific features in the sentence scoring process. The domain-general features include a *document structure* feature, which takes into account a sentence's level in terms of HTML headings, and a *content word density* feature, which computes the ratio of content words to function words. The domain-specific features include an *educational standards* feature, which uses a TF-IDF based textual similarity score between a sentence and nationally recognized educational goals from the American Association for the Advancement of Science (AAAS) Benchmarks (Project2061., 1993) and the associated National Science Education Standards (NRC, 1996), and a *gazetteer* feature, which scores sentences highly that mention many unique names from a gazetteer of named entities.

While in the past, COGENT was used primarily as a summarization system, in the current work, we evaluate its utility as a means of identifying core concepts. That is, are the top sentences selected by COGENT also the sentences describing the key science concepts in the domain?

4.2 Evaluation

We evaluate the core concept extraction module by assessing the extracted concepts against human expert annotations. We ran an annotation study where two human experts assigned "coreness" ratings to a selected set of sentences collected from digital resources in three science domains: *Plate Tectonics*, *Weather and Climate*, and *Biological Evolution*. These experts had been recruited based on their training and expertise in the selected subject domains.

First, a set of digital resources was selected from the Digital Library for Earth System Education (DLESE) [2] across the three subject domains. Then COGENT was used to extract the top 5% sentences for each domain. The experts then annotated each extracted sentence with its coreness rating on a scale of 1 to 4, 4 being the highest. Human annotation is a time-consuming process and this is why we had to limit the number of extracted sentences to a moderate 5% (which is still more than 400 sentences). 17% of the sentences were double annotated and the inter-rater reliability, measured by Spearman's rho, was 0.38. These expert ratings of sentences form the basis of our evaluation.

	Extraction %			
	0.5%	1.0%	2.5%	5.0%
Plate Tectonics	3.33	3.27	3.00	2.81
Weather and Climate	3.13	2.97	3.07	2.99
Biological Evolution	2.00	2.13	2.46	2.25

Table 1: Average coreness of sentences extracted at different percentages in each domain

Table 1 shows the average coreness assigned by the experts to sentences extracted by COGENT in each domain, for different extraction percentages. For example, if COGENT is used to extract the top 1% of sentences from all the *Plate Tectonics* resources, then the average of their coreness ratings (as assigned by the experts) is 3.27, representing a high level of coreness. This is essentially a measure of the precision of COGENT at 1% extraction. Note that we cannot calculate a measure of recall without asking experts to annotate all of the domain sentences, a time consuming task which was outside of the scope of this study.

The performance of COGENT was the best in the *Plate Tectonics* domain since the domain-aware features (e.g. the gazetteer features) used to train COGENT were selected from this domain. In the "near domain" of *Weather and Climate*, the performance is still good, but performance falls in the "far domain" of *Biological Evolution*, because of the significant differences between the training domain and the test domain. In the two latter domains, the performance of COGENT was also inconsistent in that with an increase in the extraction percentage, the average coreness increased in some cases and decreased in others. This inconsistency and overall degradation in performance in the two latter domains are indicative of the importance of introducing domain-aware features into COGENT.

It is evident from the values in Table 1 that the core concepts extraction module does a decent job,

[2] http://www.dlese.org

especially when trained with appropriate domain-aware features.

5 Sequencing core concepts

The goal of this next component is to take a set of core science concepts (sentences), as produced by the preceding module, and predict an appropriate sequence in which those concepts should be learned by the student. Some concepts serve as building blocks for other concepts, and thus it is essential to learn the basic concepts first (and address any misconceptions associated with them) before moving on to other concepts that depend on the basic concepts. For example, a student must first understand the concept of tectonic plates before they can understand the concept of a convergent plate boundary. The sequence of core concepts that results from this module will serve as input for the later module that prioritizes a student's misconceptions.

There may exist several different but reasonable concept sequences (also known as *learning paths*) – the goal of this component is to recommend at least one of these. As a first step, we focus on generating a single concept sequence that represents a general path through the learning goals, much like textbooks and curriculums do.

5.1 Models

Our model for concept sequencing is a pair-wise ordering model, that takes two concepts c_1 and c_2, and predicts whether c_1 should come before or after c_2 in the recommended learning path. Formally,

$$\text{SEQUENCE}(c_1, c_2) = \begin{cases} 0 & \text{if } c_1 < c_2 \\ 1 & \text{if } c_1 \geq c_2 \end{cases}$$

To generate a complete ordering of concepts, we construct a precedence table from these pair-wise judgments and generate a path that is consistent with these judgments.

We learn the SEQUENCE model as a supervised classifier, where a feature vector is extracted for each of the two concepts and the two feature vectors, concatenated, serve as the input to the classifier. For each word in each concept, we include the following two features:

- **local word count** - the number of times the word appeared in this concept
- **global word count** - the log of the ratio between the number of times the word occurred in the concept and the number of times it occurred in a background corpus, Gigaword (Graff, 2002)

These features are motivated by the work of Tanaka-ishii et al (2010) that showed that local and global word count features were sufficient to build a pair-wise readability classifier that achieved 90% accuracy.

For the supervised classifier, we consider naive Bayes, decision trees, and support vector machines.

5.2 Evaluation

To evaluate our concept sequencing model, we gathered learning paths from experts in high school earth science. Using the model from Section 4, we selected 30 core concepts for the domain of plate tectonics. We asked two earth science experts to each come up with two learning paths for these core concepts, with the first path following an *evidence or research based* and second path following a *traditional* learning path.

An *evidence or research based* learning path, is a pedagogy where students are encouraged to use the scientific method to learn about a phenomena, i.e they gather information by observing the phenomena, form a hypothesis, perform experiment, collect and analyze data and then interpret the data and draw conclusions that hopefully align with the current understanding about the phenomena. A teacher that uses this learning path acts as a *guide on the side*. A *traditional* learning path on the other hand, is the pedagogy where teachers are simply trying to pass on the correct information to students rather than letting the students discover the information themselves. In a classroom environment, a teacher using this learning path would be seen as the classical *sage on stage*.

We used the learning paths collected from the experts to form two test sets, one for the *evidence-based* pedagogy, and one for the *traditional* pedagogy. For each pedagogy, we asked which of all the possible pair-wise orderings our experts agreed upon. For example, if the first expert said that $A < B < C$ and the second expert said that $A < C < B$, then both experts agreed that $A < B$ and $A < C$, while they disagreed on whether $B < C$ or $C < B$. Note that we evaluate pair-wise orderings here, not a complete ranking of the concepts, because the experts did not

Pedagogy	*Pairs (%)*	$c_1 < c_2$	$c_1 \geq c_2$
Evidence	637 (68%)	48.5%	51.5%
Traditional	613 (70%)	48.5%	51.5%

Table 2: Test sets for sequencing concepts. The *Pairs* column shows how many pairs the experts agreed upon (out of a total of $30 * 29 = 870$ pairs).

	Baseline	*NaiveBayes*	*SVM*
Evidence	51.5%	60.8%	53.3%
Traditional	51.5%	56.6%	49.7%

Table 3: Accuracy result from Naive Bayes and SVM for classifying the core concepts

produce a total ordering of the concepts, only a partial tree-like ordering. The experts put the concepts in levels, with concepts in the same level having no precedence relationship, while a concept in a lower level preceded a concept in a higher level.

For our test sets, we selected only the pairs on which both experts agreed. Table 2 shows that experts agreed on 68-70% of the pair-wise orderings. Table 2 also shows the percentage of each type of pair-wise ordering ($c_1 < c_2$ vs. $c_1 \geq c_2$) present in the data. Note that even though all concepts are paired with all other concepts, because the experts do not produce complete orderings, the number of agreements for each type of ordering may not be the same. Consider the case where expert E_1 says that concepts A and B are on the same level (i.e., $A = B$) and expert E_2 says that concept A is in a lower level than concept B (i.e., $A < B$). Then for the pair (A, B), they disagree on the relation (E_1 says $A \geq B$ while E_2 says $A < B$) but for the pair (B, A) they agree on the relation (they both say $B \geq A$). As a result, the $c_1 \geq c_2$ class is slightly larger than the $c_1 < c_2$ class.

Since these data sets were small, we reserved them for testing, and trained our pair-wise classification model using a proxy task: ordering sentences by grade. In this task, the model is given two sentences s_1 and s_2, one written for middle school and written for high school, and asked to decide whether $s_1 < s_2$ (i.e. s_1 is the middle school sentence) or $s_2 < s_1$ (i.e. s_2 is the middle school sentence). We expect that a model for ordering sentences by grade should also be a reasonable model for ordering concepts for a pedagogical learning path. And importantly, getting grade ordering data automatically is easy: the Digital Library for Earth System Education (DLESE) contains a variety of earth science resources with metadata about the grade level they were written for.

To construct the training data, we searched the DLESE website for text resources that contained the words *earthquake* or *plate tectonics*. We collected 10 such resources for each of the two grade cohorts, middle school (we allowed anything K-8) and high school (we allowed anything 9+). We downloaded the webpage for each resource, and used COGENT to extract the 20 most important sentences from each. This resulted in 200 sentences for each of the two grade cohorts. To create pairs of grade-ordered sentences, we paired up middle and high school concepts both ways: middle school first (i.e. $\text{SEQUENCE}(c_m, c_h) = 0$) and high school first (i.e. $\text{SEQUENCE}(c_h, c_m) = 1$). This resulted in 40,000 grade-ordered sentence pairs for training.

We then used this proxy-task training data to train our models. We extracted 1702 unique non-stopwords from the training data, resulting in 3404 features per concept, and 6808 features per concept pair (i.e. per classification instance). On the grade-ordering task, we evaluated three models using WEKA[3], a naive Bayes model, a decision tree (J48) model, and a support vector machine (SVM) model. Using a stratified 50/50 split of the training data, we found that the naive Bayes and SVM models both achieved an accuracy of 80.2%, while the decision tree achieved only 62%. So, we selected the naive Bayes and SVM models for our real task, concept sequencing.

Table 3 shows the performance of the two models on the expert judgments of concept sequencing. We find that the naive Bayes model produces more expert-like concept sequences than would be generated by chance and also outperforms the SVM model on the concept sequencing task. For the final output of the module, we combine the pair-wise judgments into a complete concept sequence, breaking any ties in the pair-wise judgments by preferring the order of the concepts in the output of the core concept identifier.

[3] http://www.cs.waikato.ac.nz/ml/weka/

6 Identifying student misconceptions

The previous components have focused on analyzing the background knowledge – finding core concepts in the domain and selecting an appropriate learning sequence for these concepts. The current component focuses on the student essay, using the collected background knowledge to help analyze the essay and give feedback.

Given a student essay, the goal of this component is to identify which sentences in the essay are most likely to be misconceptions. The task of misconception identification is closely related to the task of textual entailment (Dagan et al., 2006), in which the goal is to predict if a hypothesis sentence, H, can be reasonably concluded given another sentence, T. In misconception identification, the goal is to predict if a student sentence can be concluded from any combination of the sentences in the domain knowledge, similar to a textual entailment task with a single H but many Ts. A student sentence that can not be concluded from the domain knowledge is likely a misconception.

6.1 Models

We developed two models for identifying student misconceptions, inspired by work in textual entailment that showed that a model that simply counts the words in H that appeared in T, after expanding the words in T using WordNet, achieves state-of-the-art performance (Shnarch et al., 2011)[4].

The **Coverage** model scores a student sentence by counting the number of its words that are also in some domain sentence. Low-scoring sentences are likely misconceptions. Formally:

$$\text{SCORE}(s) = \frac{|s \cap d|}{|s|} \qquad d = \bigcup_{s' \in D} \text{EXPAND}(s')$$

where s is a student sentence (a list of words), D is the set of domain sentences, and EXPAND performs lexical expansion on the words of a sentence.

The **Retrieval** model indexes the domain sentences with an information retrieval system (we use Lucene[5]), and scores a student sentence by querying the index and summing the scores. Formally:

$$\text{SCORE}(s) = \sum_{s' \in D} \text{SCORE}_{\text{lucene}}(s, \text{EXPAND}(s'))$$

where s, D and EXPAND are defined as before, and $\text{SCORE}_{\text{lucene}}$ is a cosine over TF-IDF vectors[6].

For both the **Coverage** and **Retrieval** models, we consider the following lexical expansion techniques for defining the EXPAND function:

- **tokens** – words in the sentence (no expansion)
- **tokens, synsets** – words in the sentence, plus all lemmas of all WordNet synsets of each word
- **tokens, synsets$_{\text{expanded}}$** – words in the sentence, plus all lemmas of all WordNet synsets of each word, plus all lemmas of derived forms, hyponyms or meroynms of the WordNet synsets
- **tokens, synsets$_{\text{expanded}\times 4}$** – words in the sentence, plus all lemmas of all WordNet synsets of each word, plus all lemmas of WordNet synsets reachable by a path of no more than 4 links through derived forms, hyponyms or meroynms

6.2 Evaluation

We evaluate the quality of our misconception identification models using data collected from the American Association for the Advancement of Science's Project 2061 Science Assessment Website[7]. This website identifies the main ideas in various topics under Life Science, Physical Science and Earth Science, and for each idea provides several sentences of description along with its individual concepts and common student misconceptions.

We used 3 topics (17 ideas, averaging 6.2 description sentences, 7.1 concept sentences and 9.9 misconception sentences each) as a development set:

CE	Cells
AM	Atoms, Molecules, and States of Matter
PT	Plate Tectonics

We used 11 topics (64 ideas, averaging 5.9 description sentences, 9.4 concept sentences and 8.6 misconception sentences each) as the test set:

[4]The paper also proposes a more elaborate probabilistic model, but shows that the "lexical coverage" model we adopt here is quite competitive both with their probabilistic model and with the top-performing systems of RTE5 and RTE6.

[5]http://lucene.apache.org

[6]See org.apache.lucene.search.Similarity javadoc for details.

[7]http://assessment.aaas.org/

Model	*MAP*	*P@1*
Randomly ordered	0.607	0.607
Coverage - tokens	0.647	0.471
Coverage - tokens, synsets	0.633	0.529
Coverage - tokens, $\text{synsets}_{\text{expanded}}$	0.650	0.471
Coverage - tokens, $\text{synsets}_{\text{expanded}\times 4}$	**0.690**	**0.706**
Retrieval - tokens	0.665	0.529
Retrieval - tokens, synsets	0.641	0.471
Retrieval - tokens, $\text{synsets}_{\text{expanded}}$	0.650	0.529
Retrieval - tokens, $\text{synsets}_{\text{expanded}\times 4}$	**0.684**	**0.647**

Table 4: Development set results for identifying misconceptions.

EN	Evolution and Natural Selection
BF	Human Body Systems
IE	Interdependence in Ecosystems
ME	Matter and Energy in Living Systems
RH	Reproduction, Genes, and Heredity
EG	Energy: Forms, Transformation, Transfer…
FM	Force and Motion
SC	Substances, Chemical Reactions…
WC	Weather and Climate: Basic Elements
CL	Weather and Climate: Seasonal Differences
WE	Weathering, Erosion, and Deposition

For the evaluation, we provide all of the idea's description sentences as the domain knowledge, and combine all of an idea's concepts and misconceptions into a "student essay"[8]. We then ask the system to rank the sentences in the essay, placing misconceptions above true concepts. Accuracy at placing misconceptions at the top of the ranked list is then measured using mean average precision (MAP) and precision at the first item (P@1).

The models were compared to a chance baseline: the expected MAP and P@1 if the concept and misconception sentences were ordered randomly. Table 4 shows that on the development set, while all models outperformed the random ordering baseline's MAP (0.607), only models with lexical expansion from 4-link WordNet chains outperformed the baseline's P@1 (0.607). The Coverage and Retrieval models using this expansion technique had comparable MAPs (0.690 vs. 0.684), but the Coverage model had a higher P@1 (0.706 vs. 0.647). These top two misconception identification models were evaluated on the test set. Table 5 shows that both models again outperformed the random ordering baseline, and the Retrieval model outperformed the Coverage model (0.644 vs. 0.603 MAP, 0.625 vs. 0.578 P@1).

[8]These "student essays" are a naive approximation of real essays, but the sentences are at least drawn from real student errors. In the future, we hope to create an evaluation corpus where real student essays have been annotated for misconceptions.

Model	*MAP*	*P@1*
Randomly ordered	0.487	0.487
Coverage - tokens, $\text{synsets}_{\text{expanded}\times 4}$	0.603	0.578
Retrieval - tokens, $\text{synsets}_{\text{expanded}\times 4}$	**0.644**	**0.625**

Table 5: Test set results for identifying misconceptions.

7 Aligning misconceptions to concepts

The goal of this component is to take the misconception sentences identified in a student essay and align them to the core science concepts identified for the domain. For example, a student misconception like *Earth's plates cannot bend* would be aligned to a science concept like *Mountains form when plate material slowly bends over time*.

7.1 Models

The model for misconception-concept alignment takes a similar approach to that of the Retrieval model for misconception identification. The alignment model applies lexical expansion to each word in a core science concept, indexes the expanded concepts with an information retrieval system, and scores each concept for its relevance to a student misconception by querying the index with the misconception and returning the index's score for that concept. Formally:

$$\text{SCORE}(c) = \text{SCORE}_{\text{lucene}}(m, \text{EXPAND}(c))$$

where m is the query misconception, c is the science concept, and EXPAND and $\text{SCORE}_{\text{lucene}}$ are defined as in the Retrieval model for misconception identification. The concept with the highest score is the concept that best aligns to the student misconception according to the model.

For lexical expansion, we consider the same definitions of EXPAND as for misconception identification: **tokens**; **tokens, synsets**; **tokens, $\text{synsets}_{\text{expanded}}$**; and **tokens, $\text{synsets}_{\text{expanded}\times 4}$**.

Model	*MAP*	*P@1*
Randomly ordered	0.276	0.276
Alignment - Tokens	0.731	0.639
Alignment - Tokens, synsets	**0.813**	**0.734**
Alignment - tokens, synsets$_{\text{expanded}}$	0.790	0.698
Alignment - Tokens, synsets$_{\text{expanded}\times 4}$	0.762	0.639

Table 6: Development set results for aligning concepts to misconceptions.

7.2 Evaluation

We again leverage the AAAS Science Assessments to evaluate the misconception-concept alignment models. In addition to identifying key science ideas, and the concepts and common misconceptions within each idea, the AAAS Science Assessments provide links between the misconceptions and the concepts. Usually there is a single concept to which each misconception is aligned, but the AAAS data aligns as many as 16 concepts to a misconception in some cases.

For the evaluation, we give the system one misconception from an idea, and the list of all concepts from that idea, and ask the system to rank the concepts[9]. If the system performs well, the concepts that are aligned to the misconception should be ranked above the other concepts. Accuracy at placing the aligned concepts at the top of the ranked list is then measured using mean average precision (MAP) and precision at the first item (P@1).

The models were compared to a chance baseline: the expected MAP and P@1 if the concept and misconception sentences were ordered randomly. Table 6 shows that on the development set, all models outperformed the random ordering baseline. Lexical expansion with tokens and synsets achieved the highest performance, 0.813 MAP and 0.734 P@1. This model was evaluated on the test set, and Table 7 shows that the model again outperformed the random ordering baseline, achieving 0.704 MAP and 0.611 P@1. Overall, these are promising results – given a student misconception, the model's first choice for a concept to address the misconception is helpful more than 60% of the time.

[9]As discussed in Section 6.2, there are on average 9.4 concepts per item. This is not too far off from the 10-20 core concepts we typically expect the tutor to extract for each domain.

Model	*MAP*	*P@1*
Randomly ordered	0.259	0.259
Alignment - Tokens, synsets	**0.704**	**0.611**

Table 7: Test set results for aligning concepts to misconceptions.

8 Recommending resources

The goal of this component is to take a set of student misconceptions, the core science concepts to which each misconception is aligned, and the pedagogical ordering of the core science concepts, and recommend digital resources (webpages) to address the most important of the misconceptions. For example, a student that believes that *water evaporates into the air only when the air is very warm* might be directed to websites about evaporation and condensation. The recommended resources are intended to help the student quickly locate the concept knowledge necessary to correct each of their misconceptions.

8.1 Models

The intuition behind our model is simple: sentences from recommended resources should contain the same or lexically related terminology as both the misconception sentences and their aligned concepts. As a first approach to this problem, we focus on the overlap between recommended sentences and the misconception sentences, and use an information retrieval approach to build a resource recommender.

First, the user gives the model a set of domain knowledge webpages, and we use an information retrieval system (Lucene) to index each sentence from each of the webpages. (Note that we index all sentences, not just core concept sentences.) Given a student misconception, we query the index and identify the source URL for each sentence that is returned. We then return the list of the recommended URLs, keeping only the first instance of each URL if duplicates exist. Formally:

$$\text{SCORE}(url) = \max_{s \in url} \text{SCORE}_{\text{lucene}}(m, s)$$

where url is a domain resource, s is a sentence from a domain resource and m is the student misconception. URLs are ranked by score and the top k URLs are returned as recommendations.

8.2 Evaluation

As a preliminary evaluation of the resource recommendation model, we obtained student misconception sentences that had been aligned to concepts in a knowledge map of plate tectonics (Ahmad, 2009). The concepts in the knowledge map were originally drawn from 37 domain webpages, thus each concept could serve as a link between a student misconception and a recommended webpage. For evaluation, we took all 11 misconceptions for a single student, where each misconception had been aligned through the concepts to on average 3.4 URLs. For each misconception, we asked the recommender model to rank the 37 domain URLs in order of their relevance to the student misconception.

We expect the final interactive essay writing system to return up to $k = 5$ resources for each misconception, so we evaluated the performance of the recommender model in terms of precision at five (P@5). That is, of the top five URLs recommended by the system, how many were also recommended by the experts? Averaging over the 11 student misconception queries, the current model achieves P@5 of 32%, an acceptable initial baseline as randomly recommending resources would achieve only P@5 of 9%.

9 Discussion

In this article, we have presented our initial steps towards an interactive essay writing system that can help students identify and remedy misconceptions in their science knowledge. The system relies on techniques drawn from a variety of areas of natural language processing research, including multi-document summarization, textual entailment and information retrieval. Each component has been evaluated independently and demonstrated promising initial performance.

A variety of challenges remain for this effort. The core concept identification system performs well on the plate tectonics domain that it was originally developed for, but poorer on more distant domains, suggesting the need for more domain-independent features. The model for sequencing science concepts pedagogically uses only the most basic of word-based features, and could potentially benefit from features drawn from other research areas such as text readability. The misconception identification and alignment models perform well on the AAAS science assessments but have not yet been evaluated on real student essays, which may require moving from lexical coverage models to more sophisticated entailment models. Finally, the recommender model considers only information about the misconception sentence (not the aligned core concept nor the pedagogical ordering of concepts) and recommends entire resources instead of directing students to specifically relevant sentences or paragraphs.

Perhaps the most important challenge for this work will be moving from evaluating the components independently to a whole-system evaluation in the context of a real essay writing task. We are currently designing a study to gather data on students using the system, from which we hope to derive information about which components are most reliable or useful to the students. This information will help guide our research to focus on improving the components that yield the greatest benefits to the students.

References

[Ahmad2009] Faisal Ahmad. 2009. *Generating conceptually personalized interactions for educational digital libraries using concept maps*. Ph.D. thesis, University of Colorado at Boulder.

[Azevedo et al.2008] Roger Azevedo, Amy Witherspoon, Arthur Graesser, Danielle McNamara, Vasile Rus, Zhiqiang Cai, Mihai Lintean, and Emily Siler. 2008. MetaTutor: An adaptive hypermedia system for training and fostering self-regulated learning about complex science topics. In *Meeting of Society for Computers in Psychology*, November.

[Dagan et al.2006] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer Berlin / Heidelberg.

[de la Chica et al.2008a] Sebastian de la Chica, Faisal Ahmad, James H. Martin, and Tamara Sumner. 2008a. Pedagogically useful extractive summaries for science education. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

[de la Chica et al.2008b] Sebastian de la Chica, Faisal Ahmad, Tamara Sumner, James H. Martin, and Kirsten Butcher. 2008b. Computational foundations for personalizing instruction with digital libraries. *International Journal on Digital Libraries*, 9(1):3–18, July.

[Graesser et al.2004] Arthur Graesser, Shulan Lu, George Jackson, Heather Mitchell, Mathew Ventura, Andrew Olney, and Max Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods*, 36:180–192.

[Graff2002] David Graff. 2002. English Gigaword. *Linguistic Data Consortium*.

[NRC1996] National Research Council NRC. 1996. *National Science Education Standards*. National Academy Press, Washington DC.

[Project2061.1993] Project2061. 1993. *Benchmarks for Science Literacy*. Oxford University Press, New York, United States.

[Radev et al.2004] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938, November.

[Shnarch et al.2011] Eyal Shnarch, Jacob Goldberger, and Ido Dagan. 2011. A probabilistic modeling framework for lexical entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 558–563, Portland, Oregon, USA, June. Association for Computational Linguistics.

[Tanaka-Ishii et al.2010] K. Tanaka-Ishii, S. Tezuka, and H. Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

[VanLehn et al.2002] Kurt VanLehn, Pamela Jordan, Carolyn Rosé, Dumisizwe Bhembe, Michael Böttner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Michael Ringenberg, Antonio Roque, Stephanie Siler, and Ramesh Srivastava. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In Stefano Cerri, Guy Gouardères, and Fábio Paraguaçu, editors, *Intelligent Tutoring Systems*, volume 2363 of *Lecture Notes in Computer Science*, pages 158–167. Springer Berlin / Heidelberg.

[Wade-Stein and Kintsch2004] David Wade-Stein and Eileen Kintsch. 2004. Summary Street: Interactive computer support for writing. *Cognition and Instruction*, 22(3):333–362.