Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Validating Problems of Understanding Extracted from Science Video Comments

Nils MALZAHN[a*], Christina SCHNEEGASS[a] & H. Ulrich HOPPE[a]
[a]*Rhine-Ruhr Institute for Applied Systeminnovation Germany*
*nm@rias-institute.eu

Abstract. The JuxtaLearn project (funded by the EU) aims at facilitating the acquisition of science concepts through the creation and sharing of videos on the part of the learners. For the specific learning targets threshold concepts are specified as key elements of knowledge. Content analysis techniques are used to extract learners' concepts manifested in textual artifacts taken from online resources like Khan Academy and YouTube comments. Deviations between student concepts and an agreed upon domain knowledge represented in an ontology can indicate problems of understanding. In this paper we particularly report on the judgement of teachers reading the validity of the problems of understanding identified by our method.

Keywords: conceptual change, network text analysis, STEM learning, video based learning

1. Introduction

The on-going European project JuxtaLearn explores the potential of fostering learning in different fields of science (or STEM) by stimulating curiosity and understanding through creative performance on the part of the students in terms of creative video making/editing and sharing/commenting activities. In order to intelligently analyze and support the sharing and commenting of videos, we have adapted and used specific techniques for the analysis of learner created textual artifacts to characterize the learners' understanding of science concepts in terms of semantic networks (Daems et al., 2014). In an initial phase of the project, the textual artifacts were video comments from existing web-based learning platforms. Although the problems of understanding that we could identify in this way were plausible, the question remains if the method identifies concepts relevant under human (teachers') judgement. In the study reported in this paper we have asked teachers to judge the identified items (i.e. problems of understanding).

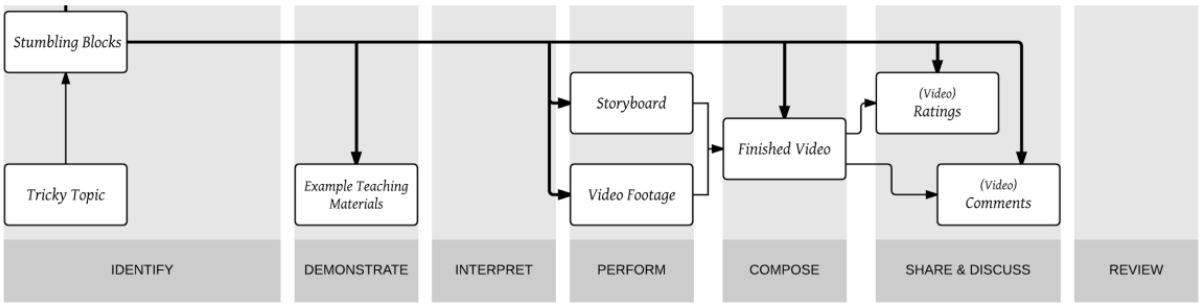

Figure 1. Juxtalearn Process

The JuxtaLearn approach can be seen as a kind of "second order inquiry leaning" in that the creative process follows an initial phase in which the learners appropriate the basic concepts of the domain. No specific assumptions about this prior knowledge building process are made. The ensuing JuxtaLearn-specific process (cf. Figure 1) comprises eight steps: (1) Identification of tricky topics, (2) demonstration of subject matter, (3) interpretation of the subject matter by the students, (4) video enactment, (5) composition of a video, (6) sharing the video with others, (7) discussion of the video and (8) review of the results. As shown in Figure 1 the tricky topics and their particular stumbling blocks feed through the whole process. From an educational design perspective, the JuxtaLearn

approach is based on the notion of "threshold concepts" (Meyer & Land, 2003) to characterize knowledge elements that enable important shifts of understanding. In the JuxtaLearn approach these are represented as so-called "tricky topics" with subordinate "stumbling blocks". Initially, the selection of tricky topics and stumbling blocks depends on the teachers' choices and are not prescribed through a normative reference list. To support the teachers defining tricky topics and stumbling blocks we automatically extract and suggest potential problems of understanding organized in an evolving problem ontology.

In JuxtaLearn, we have used external sources, such as Khan Academy[1] or YouTube[2], to develop and test our analytic methods as they provide a vast amount of videos on different STEM topics and offer the option to enter into a learning dialogue with other users (students).

2. Related Work

Learning may be seen as a process of knowledge revision and conceptual change (cf. e.g., Chi, 2008). The acquisition of new concepts may either be an extensions of the learner's pre-knowledge (called enrichment by Chi, 2008) or knowledge revisions that arise from cognitive conflicts between pre-knowledge and new phenomena or dependencies to be explained (Vosniadou, 2007).

To make the progression of understanding susceptible to inspection and reflection it is desirable to make the students' conceptual models visible for them and for their teachers. Technically this can be supported by analyzing student-generated texts (comments or notes) and transforming them into network representations. A method that provides this basic function is "Network Text Analysis" (NTA, cf. Carley, Columbus, & Landwehr, 2013). In this approach, a concept stands for a single idea, which is represented by one or more words in a network (nodes). The links representing semantic relationships between these words (edges) may differ in strength, directionality, and type based on the words' position to each other in the text. Similar to text networks, concept maps are networks in which knowledge is represented by concepts and their relationships to each other (Novak & Cañas, 2008). In the context of knowledge construction research, concept maps are often used to trace the student's knowledge development (Engelmann, Dehler, Bodemer, & Buder, 2009).

We have adopted this network-analytic perspective for the analysis of textual learning objects, following up on a suggestion by Jacobsen and Kapur (2010) to conceive learners' mental models or "ontologies" as scale-free networks. This approach has been further elaborated by Hoppe, Engler and Weinbrenner (2012) proposing a network-based model of knowledge evolution and conceptual change.

3. Approach

Our analysis of learner generated textual artefacts mainly relies on the categorical distinction between domain concepts, pedagogical and general concepts. Basic taxonomic relations should be provided for domain concepts, but simple dictionaries or lists of words are sufficient for the other two categories. Following Guarino et al's (2009) classification of ontologies we use an informal taxonomy that is rooted in the hierarchy of categories from Wikipedia (w.r.t. knowledge of the domain) and the contributions of teachers (w.r.t. knowledge about threshold concepts and stumbling blocks). In Daems et al. (2014), we show that the semantic networks generated from example sources are human-interpretable and allow for extracting potential problems of understanding.

As a source for external data we use (again) Khan Academy's video library (more than 4.300 on numerous science topics on different educational levels; Khan Academy, 2013). For our case study we used eight videos covering topics from chemistry, biology, and physics. In total, we have extracted 4476 comments from Khan Academy and YouTube for these videos using each site's particular web service. Table 1 provides an overview of the topics and amount of comments. The video length ranges from 09:50 minutes (Thermodynamics) to 19:15 minutes (Entropy) with an average of about 15 minutes.

[1] https://www.khanacademy.org

[2] https://www.youtube.com

As explained by Daems et al (2014) we found that certain general (not domain-specific) keywords are often used to phrase questions around the videos. These words (e.g. related to an explanation request) may indicate specific problems of understanding that may occur in combination with one or more domain concepts. These "signal concepts" indicate a specific problem related to one domain concept or related to two domain concepts. Therefore, we introduce two types of signal concepts: unary and binary. In this study we focus on binary signal concepts referencing two domain concepts in combination or inter-relation. A typical example for the binary type is "difference_between", which may indicate that the author has a cognitive conflict (Vosniadou, 2007) concerning the difference between the two domain concepts.

Table 1: Topics and added number of extracted comments on STEM videos at Khan Academy's website and the corresponding YouTube videos.

Subject	Topic	# comments
Chemistry	Covalent and Metallic Bonds	855
	Elements and Atoms	1934
	Le Chatelier's Principle	196
Biology	Chromosomes	487
	Mitosis	748
Physics	Entropy	39
	First Law of Thermodynamics	94
	Thermodynamics	123

To extract the network from the text we use a text window running over the text source (i.e. a pre-processed comment) to detect the co-occurrences in this context. For the results presented below we used a window size (7) guaranteeing that the signal concept and the domain concept(s) share a context with at most five words in-between.

Figure 2a shows the "cloud" of concept pairs around the concept "difference_between" generated from 855 comments of the video "Covalent and Metallic Bonds". The red node represents the signal concept this network focuses on while the blue ones are helper nodes ("combination nodes") that clearly identify the corresponding nodes for each of the combinations (see also Daems et al. 2014). The size of the combination nodes represents their frequency of occurrence throughout all comments. Green nodes represent the domain concepts themselves.

As a first step towards a set of potential problems of understanding to be presented to a teacher, all concepts that do not meet a survival threshold are dropped from the set of candidates. In our study we used a minimum occurrence value of 2 for the concept nodes. All combination nodes connected to deleted nodes were removed as well. This leads to a network like it is shown in Figure 2a. To deal with very dense networks, the minimal occurrence frequency may be adapted.

For further visual exploration the number of links may be decreased further by dropping the links from the combination nodes (blue) to the concept nodes (green) as well. The concept pairs are just labels for the combination nodes then as shown in 2b. Like in figure 2b the combination nodes as well as the links are sized in proportion to the pair's occurrence frequency in the network. From this representation a ranked list of concepts is derived to be presented to teachers. This list can then be added to the domain taxonomy and proposed as stumbling blocks during the "Identify" stage of the JuxtaLearn learning process. For the sake of our analysis, a complete set of networks was created addressing all signal concepts as well as all topics presented in Table 1.

While Daems et al. (2014) have already shown that our algorithmic approach generates plausible results, our next step is to show the validity of this approach by comparing the results of our algorithm in terms of the relevance of the proposed problems of understanding as perceived by experienced teachers.

To judge the quality of our automatic analysis we defined a set of constructs that allows deciding if a teacher might accept our suggestions. Therefore, we decided to use (1) "Plausibility" (is it reasonable to assume that there might be a problem?), (2) "Frequency" (how often is this problem to be expected to occur?), and (3) "Relevance" (Is it important to solve this issue of understanding / to be able to make this distinction?). Especially the later one is a very important trait characterizing threshold concept according to Meyer & Land (2003).

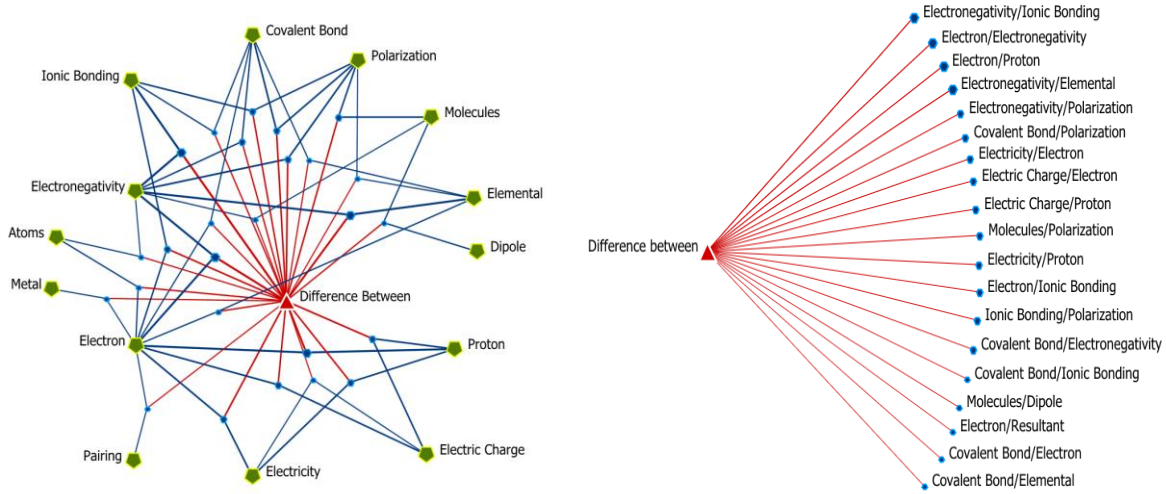

Figure 2a/b. Network on the topic of "Covalent and Metallic Bonds" with the signal concept "difference_between" (red), combination nodes (blue) and domain concepts (green).

4. Evaluation

To evaluate the results of our approach we confronted high school teachers with the results from the automatic extraction. Their expertise and knowledge and moreover their long-term experience with the learning capabilities of their students is used to judge the reliability of the analysis results.

4.1 Hypotheses

As aforementioned our analysis is supposed to reveal the students' lack of understanding. The analysis results are compared to the experienced teachers' estimation along the following hypothesis:

Combination nodes with a high support value represent concepts (or concept pairs) that challenge the student's understanding in the opinion of teachers. Support values are derived from the occurrence value of the corresponding nodes.

4.2 Experiment design

We have asked six German secondary school teachers (2 females), two for each of the STEM subjects (see Table 1), to participate in our study. The teachers have at least four years of work experience in all age groups. We used a questionnaire containing concept pairs from the binary signal concept networks "difference_between" and "relation". The selected signal concepts are supposed to uncover misunderstandings in differentiating and comprehending the context among each of the two concepts. We presented only those concept pairs to the teachers that met a threshold of at least two occurrences in the set of all video comments of a particular video.

The questionnaire comprised three questions each addressing one of the measures defined above: (1) the plausibility of the relation between the two a concept according to the teachers' expertise, (2) the occurrence frequency in class and (3) their relevance for the subject matter. Each question can be answered on a 5-point Likert-scale ranging from 0 (not at all) to 4 (very plausible/ frequent/ relevant). To avoid arbitrary answers the teachers may state that a presented concept is not part of their current or past curricula. Concepts excluded by a teacher were not further considered.

The teachers were asked to answer the questionnaire containing between five and twelve concept pairs depending on the extent of the corresponding network. The questionnaire included both concept pairs with a high and a low occurrence frequency to cover common as well as less common problems of understanding. All items were translated from the English to German. Concept pairs containing terms which are hardly to translate were dropped as they might influence the results. After finishing the questionnaire the teachers were given the chance for open comments.

4.3 Results

Table 2 presents excerpts from the results from the questionnaire. These examples show the teachers' rating for all concept pairs in the subject chemistry linked by the signal concepts "difference_between" and "relation". The corresponding videos used for the analysis dealt with the topics of "Covalent and Metallic Bonds" and "Elements and Atoms" and had 2789 comments in total on YouTube and Khan Academy.

Table 2: Evaluation of the chemistry topic "Covalent and Metallic Bonds"

Concept pair	Construct	Teacher A	Teacher B	Automated Analysis (occurences)
Electronegativity – Ionic Bonding	Plausibility	4	4	4 (16)
	Frequency	3	4	
	Relevance	2	4	
Elektronegativity – Polarization	Plausibility	4	2	3 (11)
	Frequency	4	3	
	Relevance	3	4	
Electric Charge – Electron	Plausibility	3	4	3 (10)
	Frequency	3	4	
	Relevance	4	4	
Electron – Ionic Bonding	Plausibility	3	4	3 (9)
	Frequency	2	4	
	Relevance	4	4	
Ionic Bonding – Polarization	Plausibility	2	3	1 (3)
	Frequency	2	2	
	Relevance	3	3	
Covalent Bonding – Ionic Bonding	Plausibility	2	2	2 (7)
	Frequency	3	4	
	Relevance	4	4	
Molecule – Dipole	Plausibility	3	3	2 (7)
	Frequency	3	3	
	Relevance	3	4	

On average all concept pairs were rated as very plausible, frequent and relevant ($M_{Plausibility}$ = 3; $M_{Freqeuncy}$ = 3; $M_{Relevance}$ = 4). All but three concept pairs (out of 26 in total) across the three subjects were rated (very) relevant for the subject matter by at least one of the two teachers. In addition, all but two pairs were rated (very) plausible and all but four pairs confirmed to model a (very) frequently occurring problem in class by at least one of the teachers.

To compare the results of the teachers not only on a boolean level (agree/disagree), we transformed the occurrence frequency calculated by our algorithm to the same scale as the one used by the teachers (see Table 2). As a reference figure we used the 2nd most occurring concept as we found that the top most concept is often an outlier compared to the rest. Based on the resulting relative frequencies, we used 0,25 steps to determine the rank of an algorithm result (<0,25:=0; <0,5:=1;<0,75:=3; >=0,75:=4).

We used Spearman's rank correlation coefficient to check for dependencies between these three constructs. It turned out that there is a strong correlation between the "plausibility" and "frequency" and a moderate correlation between "frequency" and "relevance". This supports our assumption that the teachers (like our algorithm) use frequency as an indicator for relevance and plausibility.

To assess the level of agreement among the teachers we had a look at the inter rater reliability. We applied Kendalls-W test to the questionnaire data. We found that significant results for all but the physics ratings, but the Kendalls-W values only range from poor to moderate agreement ($W_{Biology}$ = .272; $W_{Chemistry}$ = .298; $W_{Physics}$ = .067).

5. Conclusion

Our study has brought forth mixed results: None of the proposed problems of understanding were rejected by both experts, but the experts also did not show a strong agreement among themselves. Moreover, comments by the teachers revealed that they had problems judging the "difference" statement for closely related and overlapping concepts, which is one possible explanation for their low agreement. In addition, they stress that the relevance of a distinction strongly depends on the

educational level. In the lower grades the distinction may be less relevant than in higher classes. A potential difference in teaching experience may explain their disagreement.

Overall, relevance is judged more positively than the other criteria, which supports our initial claim that the algorithm sieves relevant problems of understanding. The judgements on frequency and plausibility differ from relevance and are quite highly correlated to each other. We assume that these criteria reflect the constraints that are related to the curricular context (i.e. normative) and to the actual experience (i.e. of practical nature).

Looking at the limits of our approach we have found that we need a considerable amount (>100) of comments to detect reasonable concept pairs. Otherwise the support for each concept pair is too low to distinguish random combinations from meaningful ones. Thus, although our proposed method is a valuable approach for huge active learning communities (e.g. MOOCs) or Open Educational Resource repositories, it is not expected to work well with small scale courses like normal class room teaching. Nonetheless, it can produce results from the aforementioned resource repositories to guide the individual small scale teaching as well.

Acknowledgement: This work was partially funded by the European Union (EU) in the context of the JuxtaLearn project under the ICT theme of FP 7. This document does not represent the opinion of the EU, and the European Union EU is not responsible for any use that might be made of its content.

References

Carley, K. M., Columbus, D., & Landwehr, P. (2013). AutoMap User's Guide 2013 (Technical Report No. CMU-ISR-13-105) (pp. 1–219). Pittsburgh: Carnegie Mellon University, Institute for Software Research. Retrieved from: www.casos.cs.cmu.edu/publications/papers/CMU-ISR-13-105.pdf

Chi, M. (2008). Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift. In S. Vosniadou (Ed.), International Handbook of Research on Conceptual Change (pp.61–82). New York, Oxon: Routledge.

Daems, O., Erkens, M., Malzahn, N., und Hoppe, H. U. (2014). Using content analysis and domain ontologies to check learners' understanding of science concepts. Journal of Computers in Education, 1-19.

Engelmann, T., Dehler, J., Bodemer, D., & Buder, J. (2009). Knowledge awareness in CSCL: A psychological perspective. Computers in Human Behavior, 25(4), 949–960.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an Ontology?. In Handbook on Ontologies (pp. 1-17). Springer Berlin Heidelberg.

Hatano, G., & Inagaki, K. (2002). Young Children's Thinking about the Biological World. New York: Psychology Press.

Hoppe, H. U., Engler, J., & Weinbrenner, S. (2012). The Impact of Structural Characteristics of Concept Maps on Automatic Quality Measurement. In J. van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences (ICLS 2012) (pp. 291–298). Sydney, Australia.

Jacobsen, M. J., & Kapur, M. (2010). Ontologies as Scale Free Networks: Implications for Theories of Conceptual Change. In Proceedings of the 9th International Conf. of the Learning Sciences (ICLS 2010), 193-194.

Meyer, J. H. F. & Land, R. (2003) Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practicing within the disciplines. IN RUST, C. (Ed.) Improving student learning theory and practice ten years on. Oxford, Oxford Centre for Staff and Learning Development (OCSLD).

Novak, J. D., & Cañas, A. J. (2008). The Theory Underlying Concept Maps and How to Construct and Use Them (Technical Report IHMC CmapTools 2006-01 Rev 01-2008 No. 2). Pensacola: Florida Institute for Human and Machine Cognition.
Retrieved from: http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlying ConceptMaps.pdf

Vosniadou, S. (2007). Conceptual Change and Education. Human Development, 50(1), 47–54.