Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education.

Using Content Analysis to Check Learners' Understanding of Science Concepts

Oliver DAEMS[a], Melanie ERKENS[a], Nils MALZAHN[a] & H. Ulrich HOPPE[b]
[a]*RIAS Institute, Duisburg, Germany*
[b]*COLLIDE Research Group, University of Duisburg-Essen, Germany*
*hoppe@collide.info

Abstract: The ongoing EU project JuxtaLearn 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 and to contrast these with the anticipated domain concepts (represented as ontology). Deviations between student concepts and the ontology can indicate problems of understanding, which may trigger a revision of the original curriculum. In two studies we explore the potential of (semi-)automated artifact analysis to identify and characterize the students' comprehension problems around knowledge artifacts. In the first study, protocols from a "flipped classroom" style teacher-student workshop are analyzed. The second study analyses comments to videos from educational video platforms. Here, we have applied text analysis methods to identify potential problems of understanding. Also, we have used "signal words" and their relations to domain concepts to highlight potential information needs and problems of understanding.

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

1. Introduction

The on-going European project JuxtaLearn aims at fostering learning and curiosity in different fields of science (or STEM) by combining or "juxtaposing" the understanding of domain concepts with performing. Concretely, the students' performance is substantiated in the form of creative video making and editing activities. We see this way of learning by performing and presenting as a variant of Papert's "constructionism" (Papert & Harel, 1991) and as similar to learning by teaching (Gartner, Kohler, & Riessmann, 1971). In this context, we are interested in studying the role of video as a medium for learning in different (including passive) forms of usage.

The design of learning activities in JuxtaLearn is guided by previously identified threshold concepts (Meyer & Land, 2003). Threshold concepts are the basis for reinforcing deeper understanding and further creative production through scaffold reflections focused on essential elements. To identify such concepts and to explore how they are understood and appropriated by teachers and students, a series of face-to-face workshops has been conducted. Following an initial workshop with science teachers, a second workshop also involved a group of six A-level students. The workshop was structured employing a role reversal with students exposing central ideas from the domains as a first step. This enabled the teachers to elicit a deeper understanding of the gaps in the students' knowledge. In our study, this was the first target of applying Learning Analytics techniques to extract structured representations of the underlying conceptual relations.

Since at this point of development the project had not yet produced collections of student-created videos, we have also tried to identify processes of understanding around videos by analyzing existing web-based learning communities, namely from Khan Academy. Since videos from Khan Academy are hosted and therefore also available on YouTube as a less "educationally guided" environment, we were able to compare comments from both contexts. We only analyzed questions and answers, not the videos nor their contents.

We were particularly interested in extracting information from these texts to shed light on the following aspects:

- Associations of concepts (which may be adequate or inadequate from a scientific point of view);

- Concepts that are frequently addressed in questions as indicators of possible origins of problems of understanding;
- Associations between concepts often used in answers as indicators for missing relations in the original understanding

We used these indicators to infer possible misconceptions and "stumbling blocks". As known from classical learner modeling (Wenger, 1987), we had to distinguish between missing knowledge about concepts and/or relations and misconceptions as often idiosyncratic constructions of incorrect knowledge. An example of a misconception (beyond missing knowledge) would be an in-correct association between two or more concepts.

2. A Network Perspective on Conceptual Models and Conceptual Change

2.1 Application of text mining techniques on learning data

In various scenarios of learning, knowledge building and knowledge production, humans externalize their knowledge in terms of "knowledge artifacts", which are often represented in the form of texts and thus susceptible to being analyzed by text mining (Heyer, Quasthoff, & Wittig, 2006). Content analysis, as a form of artifact analysis, can be used to reduce qualitative textual data into clusters of conceptual categories aiming to unfold patterns and relationships of meaning (Julien, 2008). Although several (semi-)automated methods can be used to detect these patterns from content, e.g. statistical methods based on the Vector Space Model (VSM) or Latent Dirichlet Allocation (Blei, 2012) as a probabilistic method, these methods are barely used on learning data, so far (He, 2013). Content analysis in the context of learning data has primarily been used for clustering resources, e.g. the grouping of e-learning resources according to their similarity (e.g. Hung, 2012; Tane, Schmitz, & Stumme, 2004). Sherin (2012), however, found that even without using semantic background knowledge, a Vector Space Model (VSM) based clustering of spoken word transcripts is an adequate instrument to identify student's concepts and the dynamics of their mental constructs. Additionally, He (2013) provided evidence that similar techniques are suitable for grouping learners' main topics in student-to-teacher online questions and peer-to-peer chat messages related to online video learning lessons.

The aforementioned methods are based on the "bag of word" model, in which the given order of words in a text is of no relevance to the analysis (Blei, 2012). A method that takes the words' positioning into account is the Network Text Analysis (NTA). NTA is a text mining method, which is based upon the assumption that knowledge can be modeled as a network of concepts (Carley, Columbus, & Landwehr, 2013). Against this background, a concept is a single idea, which is represented by one or more words in a network (nodes). The links representing semantic relationships between these words (edges) are differing in strength, directionality and type based on the words' position to each other in the text (Carley, Columbus, et al., 2013). The union of all relations builds the semantic network (Carley, Columbus, et al., 2013), similar to the relational network of a concept map. Similar to text networks, concept maps are networks in which knowledge is represented by concepts and their relationships to each other. They differ from text-based semantic networks inasmuch as they are arranged hierarchically with the ontological root concepts at the top (Novak & Cañas, 2008). In the context of knowledge construction research, concept maps are often used to trace the student's knowledge development (Engelmann & Hesse, 2010; Engelmann, Dehler, Bodemer, & Buder, 2009; Schreiber & Engelmann, 2010).

2.2 Representation of conceptual models as text networks

Jacobsen and Kapur (2010) have suggested to conceive learners' mental models or "ontologies" as scale-free networks, which would allow to apply known characteristics of such networks to theories of conceptual change. According to Barabási (2009) the evolution of scale-free networks can be explained by the mechanisms of "preferential attachment". Applied to learners' ontologies, preferential attachment means that newly learned concepts are most frequently associated or linked to those concepts that are already more densely connected than others. From this, Jacobsen and Kapur (2010) conclude that such "hubs" i.e. nodes with a relatively high degree centrality, represent root categories of knowledge domains. Hoppe, Engler and Weinbrenner (2012) support this theory in a study in which the volunteers had to create concept maps on the subject of global warming. This study clearly showed a scale-free nature of the maps in terms of an inverse power law degree distribution (a known structural characteristic of scale-free networks). This implies that there are more hubs than to be expected in a randomly connected network. Also, Hoppe et al. (2012) could show that certain known graph-theoretical structural measures correlate with quality judgments of these maps by independent experts. Interestingly, the "density" measure is negatively correlated with the criterion of map "completeness", which is significant. Again the scale-free model provides a clear explanation: In a growing scale-free network the density is anti-proportional to the size of the network, i.e. the smallest networks will show the highest density. Based on this characterization, the authors hypothesize that newly appearing hubs represent 'hot spots of conceptual change' (Hoppe et al., 2012, p 297), whereby this change describes a restructuring process, in which learners revise their false beliefs and misconceptions on the relational or ontological level (Chi, 2008). If the number of edges around a node is suddenly reduced, this may indicate a qualitative change of understanding or a paradigm shift (Hoppe et al., 2012). Viewing concept maps as networks allows for applying a variety of techniques known from Social Network Analysis (SNA - cf. Wasserman & Faust, 1994). As an example, the betweenness centrality measure is suggested as a possible indicator to identify "bridge concepts" that link different knowledge domains (Hoppe et al., 2012). It is important to note that the results of NTA (see above) are also networks that can be further analyzed in the same way as concept maps. This would also allow the comparison of textual input (e.g. from student essays, wikis etc.) with concept maps.

Our basic idea and approach is to use content analysis techniques to generate network representations from knowledge artifacts originally created by students or experts and to apply structural and differential (comparative) measures to these representations in order to detect similarities or mismatches. In this approach, expert maps or ontologies can be used as "normative" references for comparison, e.g. to indicate deviations from standard domain knowledge and possible misconceptions. Regarding the evolution of maps, certain structural features and anomalies can also be detected.

3. Text-based Content Analysis: Method and First Results

3.1 Network Text Analysis

According to Carley and colleagues (2013), the NTA workflow consists of three main steps: (1) data selection and extraction, (2) text pre-processing and (3) network analysis. We have applied this technique using the AutoMap/ORA toolset for NTA (Carley, Columbus, et al., 2013; Carley, Pfeffer, Reminga, Storrick, & Columbus, 2013):

As a first step, the data of interest will be selected, depending on the terms of reference, for instance the selection of comments belonging to a certain person or video. In the second step, the pre-processing functions are intended to prepare the textual data for subsequent analyses. Unneeded and unwanted concepts will be reduced through simple text cleaning functions such as the removal of extra spaces. Furthermore, this step serves to apply a) a stemming for reducing words to their root stem by removing suffixes from words, b) a delete list, which is required for the removal of non-relevant stop words (articles, auxiliary verbs etc.), and c) a manually generated thesaurus, used for replacing synonym concepts with the more standard form, for combining n-grams and to correct spelling errors. Next step of pre-processing is the identification and classification of concepts. Relevant concepts will be detected by analyzing the words' frequencies based on the following principle: Words and n-grams that appear more than x-times are considered as relevant and will be included into further analyses, whereby x depends on the size of the corpus. The classification is done by the determination of categories based on the words appearing in a concept list that includes the frequency of all words and can be reduced to the most important key words using a threshold defined by the researcher. After specifying the categories, every single concept will be assigned to one of them. Therefore, an ontology-based meta-thesaurus will be created, which later is used for generating the network. As a result of the processes described before, multimodal networks will be created, whereby the modality of the network depends on the number of categories that can be identified by the researcher.

This analysis process has been applied on the transcripts of an initial role reversal workshop surrounding the STEM topics "moles" in chemistry, "potential difference" in physics and "cells" in biology. In each lesson, two students with different school marks (both excellent / mixed / both middle) taught two teachers. The transcripts were analyzed with the method described above.

3.2 First results

As a result of this analysis process, a multimodal network of categorized concepts was generated and visualized using ORA. This network comprises the following concept categories: pedagogical concepts, domain concepts, general concepts (i.e. concepts that are neither domain specific nor pedagogic concepts), tools, and actors. We have declared actors and domain concepts as the most relevant categories. These categories are represented in a meta-thesaurus. Within this thesaurus, the actor category represents all acting persons in the lessons; teachers have been labeled as T1 to T6, the researcher staff as R1 to R4 and students as S1 to S6. The domain concept category represents discipline-specific topics associated with the lesson subjects.

The number of connections between one actor and surrounding topics (also called "degree") indicates the thematic richness of an actor's contributions. Regarding the whole semantic network, the degrees reflect that both physics students (S5 and S6) as well as one of the biology students (S1), who score high in their school marks, have also a higher total degree in the network than the other students with middle school marks. Figure 1 depicts an excerpt of the net, a two-mode network of actors and domain concepts, which contains 16 actor nodes (blue square nodes) and 71 domain concept nodes (pentagonal). The excerpt only shows 6 student actors as well as the connected domain concepts (all in blue) and the domain concepts surrounding the connected domain concepts (in grey). Additionally, every subject area and corresponding sub-activity in the workshop (chemistry, biology and physics) forms a cohesive cluster in the overall network, which are possibly interconnected by "bridge concepts".

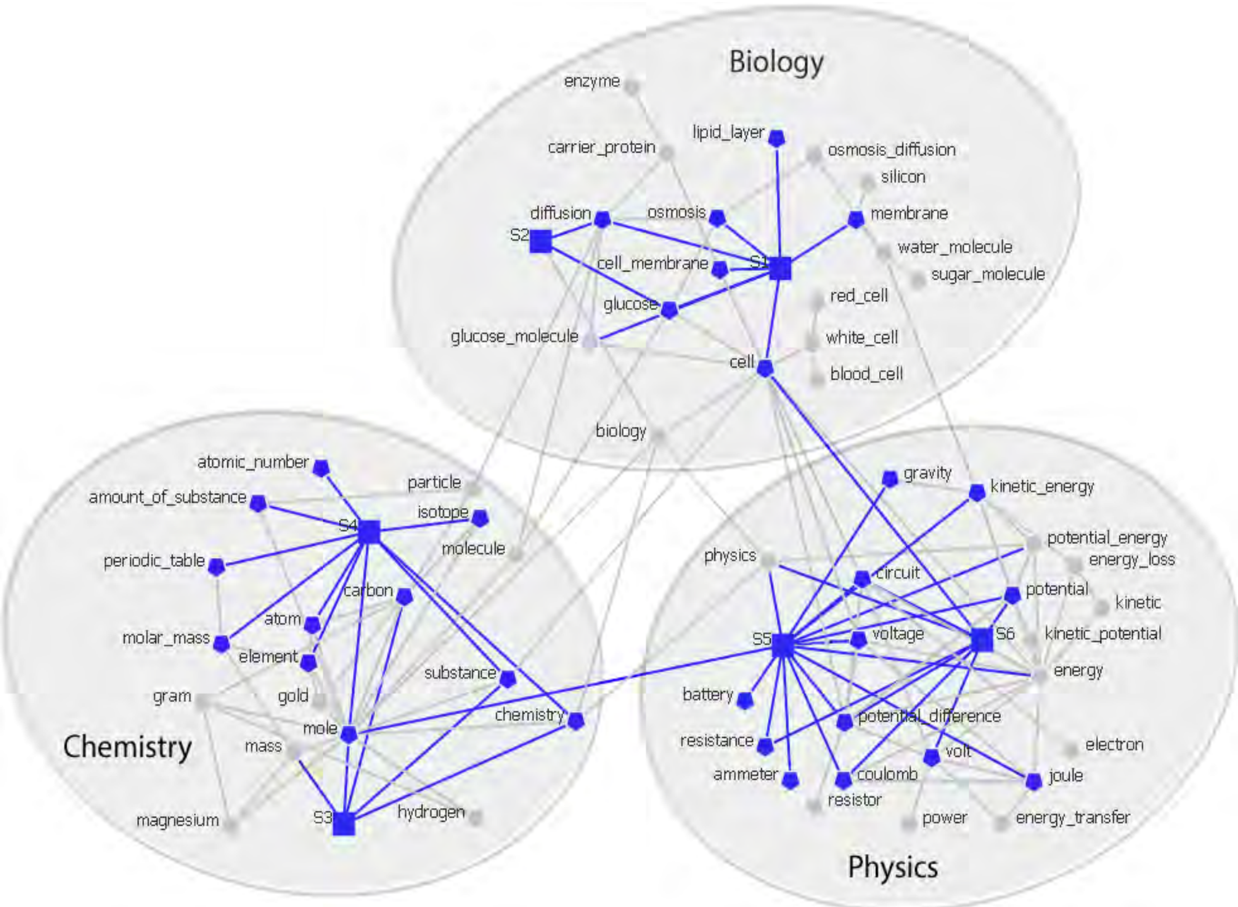

Figure 1. Partial view of a network generated from a teacher-student workshop, depicting connections of students (actors S1-S6) to domain concepts in blue (two-mode network) and relations between domain concepts in grey

Other relevant structural properties of the extracted network are nodes that represent hubs (high in total degree centrality) or nodes that bridge over between other concepts or between areas of discourse, here: the domains (high in betweenness centrality). Figure 2 shows the top six knowledge items ranked according to their total degree centrality and betweenness centrality in the uni-modal domain concept x domain concept network.

Concerning total degree centrality, we could find the three main topics of the workshops "mole", "potential difference" and "cell" within the top 6 ranking. Furthermore, the nodes high in betweenness centrality seem to bridge over different ontological areas. Mole, for instance, builds a fundamental connection between the physics related cluster and the chemistry related cluster. Furthermore, cell, second highest in betweenness centrality, also connects the three discipline clusters with each other.

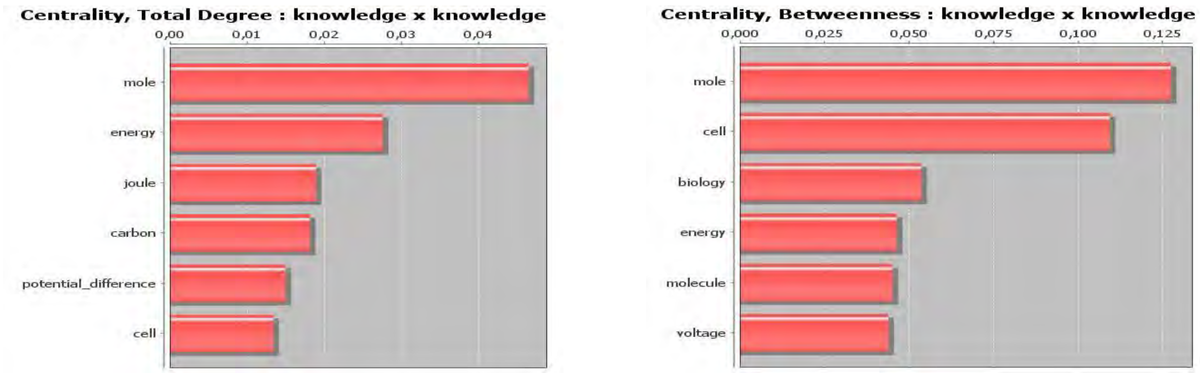

Figure 2. Domain concepts with the highest betweenness centralities (on the right) and highest total degree (on the left) in a domain concept x domain concept network.

Based on this first example, we claim that text networks can not only represent and characterize the specific foci of the JuxtaLearn workshops, but also provide ontological information, which is related to the student's conceptualization of specific science topics.

Furthermore, NTA seems to be an appropriate method to analyze student generated textual content on the JuxtaLearn online platform to learn more about their conceptual models.

4. Analysis of Video Comments

4.1 Content Analysis

Since the NTA workflow described above was conducted on offline material and suitable data from the online platform was unavailable at this developmental stage of the JuxtaLearn project, we focused on material found outside our project for further research. The approach to content analysis presented here focuses on the attention given to a specific topic in an online discussion around learning materials, which in this case are represented by online discussions composed of comments in discussions around educational videos. This approach aims to extend the NTA workflow by focusing on the content of single comments instead of the complete discussion thread.

We used a quantitative approach implementing regular expressions based matching with the concept lists constructed during the NTA workflow and the preprocessed text base which was generated during this workflow. In principal this generates a sparsely populated feature vector for each comment based on the occurrence frequency of concepts from each list. We use these vectors to construct a measure we call "semantic richness" which tries to quantify the relation between the domain specific concepts and the general concepts occurring in each comment. This measure follows the assumption that a comment that contains more domain concepts than general concepts compared to a different comment which uses a lesser ratio of domain and general concepts is more "on topic" thus semantically richer. During our analysis of external data we derived different methods of evaluating this ratio and the following will present our results.

4.2 Khan Academy and YouTube Videos

As a starting point for our content analysis we have chosen Khan Academy's learning videos and the accompanying discussion in the comments section. The Khan Academy website provides a significant amount of videos on different STEM topics and offers the option to enter into a learning dialogue with other students, a scenario which is similar to the one envisioned for the JuxtaLearn platform. The website offers message boxes below the videos to enter this dialogue through a scaffolded question/answer construct rather than an open comment section. The videos themselves are hosted on YouTube EDU, which enables users to comment on the same videos without assistive scaffolding. The library covers science topics such as biology, chemistry and physics on different levels ranging from junior high school to university and holds more than 4.300 videos with an average length of 10 minutes (Khan Academy, 2013). For this case study we used three videos with the following titles "The Mole and Avogadro's Number", "Diffusion and Osmosis" and "Voltage-Difference between electrical potential (voltage) and electrical potential energy". In total, we have extracted 1.284 comments through Khan Academy's web service. These textual artifacts have been used as a sample for our analysis. Khan Academy provides the aforementioned scaffolded discussion elements and encourages discussion on the topic of the video, which means we can assume that artifacts extracted from the discussion will be about the topic of the video as well. Table 1 gives an overview of the amount of artifacts used in our study. It shows the number of questions and answers for each area the video topic can be assigned to, as well as their sum (# of comments) for easy comparison with other sources.

YouTube EDU is a subsection of YouTube that focuses on educational videos. The idea behind it is to provide everything related to education, spanning from short lessons for supplementing school learning to full courses from universities and other professional material from educators around the globe (YouTube EDU, 2014). As mentioned above, this includes videos from the Khan Academy. For a comparison with the Khan comments, we extracted video comments from YouTube EDU (using the appropriate Google API) for the same three videos. This resulted in another 1.201 comments distributed among the videos as shown in table 2. YouTube EDU does not provide any scaffolding for discussion resulting in a mixture of comments and smaller discussion compared to Khan Academy's strict question/answer construct.

Table 1: Type and number of extracted comments on STEM videos at Khan Academy's website.

Type of artifact	Biology	Chemistry	Physics
# questions	184	279	70
# answers	312	362	77
# comments	496	641	147

Table 2: Subject and number of extracted comments on STEM videos from YouTube EDU.

Subject of artifact	Biology	Chemistry	Physics
# comments	487	628	86

4.3 Data Sampling and First Observations

The aforementioned strict construct of questions/answers seems to be enforced by a strong moderation through the Khan Academy staff, because the discussion artifacts have a very low amount of noise (e.g. spam comments). This is reflected by the high amount of very short and poorly written comments extracted from YouTube EDU. Additionally, the artifacts extracted from Khan Academy seem to focus on the video's topic rather than on the video itself, while the opposite is true for comments from YouTube EDU, which mostly contain short comments that focus on e.g. the quality of the video. This observation can be confirmed by analyzing the content based on the list of domain and other concepts we constructed during the NTA workflow. Employing a quantitative approach we counted the occurrences of general and domain concepts in artifacts from both platforms, representing these values as feature vectors for general and domain concepts. Table 3 shows the results from this approach.

Table 3: Results from statistical computations for the "semantic richness" measure.

Description	**Value**			
Source from which the basis was extracted (N represents the number of comments extracted)	**Khan Academy (N=1284) M**	**SD**	**YouTube (N=1201) M**	**SD**
sum of general and domain concepts (length of comment)	23.73	28.04	9.78	11.03
general concepts: with duplicates | only once	9.89 | 7.62	11.88 | 7.43	6.07 | 5.49	6.11 | 5.12
domain concepts: with duplicates | only once	13.84 | 7.26	17.58 | 5.60	3.71 | 2.51	6.27 | 3.45
1. measure: domain concepts / length	0.54 | 0.38	0.23 | 0.20	0.26 | 0.22	0.26 | 0.22
2. measure: domain concepts / (general concepts + 1)	1.46 | 1.08	1.48 | 1.02	0.50 | 0.40	0.88 | 0.63

The two combinatory measures were introduced to create a value describing the semantic richness. We introduced two different formulas for both versions of feature vectors (those using all occurrences therefore with duplicates / those using only the first occurrence), the first normalizes the value to a range of 0-1 but favors shorter comments as a comment containing only one domain concept and no general concepts will have a semantic richness of 1. The second favors longer comments but the value is uncapped, meaning it can grow infinitely when only domain concepts are found (the +1 is needed to avoid division by zero).

The numbers clearly indicate both longer and "semantically richer" artifacts on Khan Academy, which might be a result of the platform's focus on provision of educational content and discussion or simply the suspected strong moderation.

5. Introducing "Signal Concepts"

5.1 Idea

While scrutinizing the list of most frequent general concepts in an exploratory way, we found a subcategory of words within the comments (such as "help", "explanation", "difference_between") that seem to refer to students' problems of understanding. We decided to select the most frequent of these "signal words" in order to take a closer look at their relationships to adjacent domain concepts in the text network. These signal concepts occur in combination with domain concepts and indicate or "signal" a specific relationship either between the author and a domain concept or between two domain concepts. This distinction is reflected by using two types of signal concepts: unary and binary. Unary signal concepts reference only one domain concept and therefore represent the signal concepts that reference the author and one domain concept, while binary signal concepts reference two domain concepts. An example for a signal concept that defines a relationship between the author and a domain concept is "help_needed", which may indicate a problem the author has with the connected domain concept, while an example for the latter is "difference_between" that may indicate that the author thinks or inquires about a difference between two domain concepts. Signal concepts provide a resource for teachers interested in learning about potential problems their students might have. Thus, our approach indicates a topic or a domain concept that may be worth focusing on in a future lesson.

Initially we used our NTA workflow to single out signal concepts and visualize this part of the network, Figure 3 shows an example network based on "dont_understand", "definition" and "difference_between". This network highlights the inherent problem with analyzing signal concepts through our NTA workflow as there are several connections between the signal concept "difference_between" and different domain concepts without the possibility of identifying their original context. The missing original context means there is no way to identify which of these domain concepts were actually referenced using "difference_between", since the NTA workflow aggregates all occurrences of a signal concept into a single node. This node has connections to all domain concepts it was attributed to in the complete set of comments analyzed but lacks the original context of each individual comment.

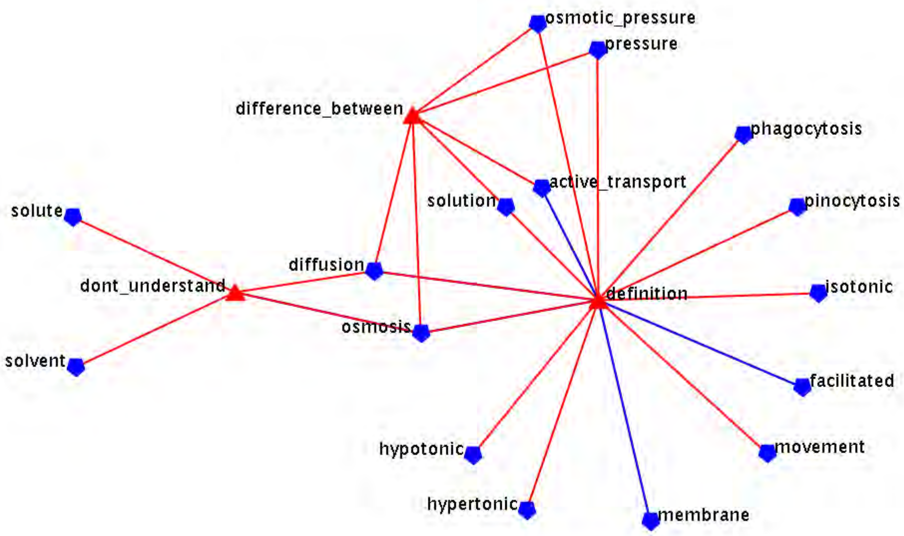

Figure 3. Network showing the neighborhood of "difference_between".

5.2 Approach

The basic idea behind our refined "signal concept" approach is to be able to correctly identify the original context and avoid the combinatorial blending seen in Figure 3. For this purpose, we generate a new network that is able to visualize the connections between the signal concept and the referenced domain concept(s) separately for each co-occurrence. Here, we did not search for co-occurrences in entire comments but in segments (windows) with a size of seven words in the pre-processed text, i.e. the signal concept and the domain concept(s) can have at most five words between them.

In a further refinement of the method we used a matching algorithm with predefined patterns of domain and signal concepts specified in the form of regular expressions. This algorithm transforms the preprocessed text into a compact string representation by replacing each word with a single letter symbol. We used six letters (G, D, S, T, M and O) which represent general concepts (G), domain concepts (D), unary signal concepts (S), binary signal concepts (T), binary signal concepts that need to be in the middle of two domain concepts (M) and other (O). The patterns we used for the matching on these string representations are shown in Table 5 and can be broken down into two archetypes, for unary and binary signal concepts. The rules reflect different orders of occurrence in the text for these two types of signal concepts.

Table 5: Patterns used for digging up signal concepts from artifacts.

Pattern	Description
(S)[OG]{0,5}(D)	Matches a unary signal (S) and a domain concept with 0-5 general or other concepts between them.
(D)[OG]{0,5}([MT])[OG]{0,5}(D)	Matches a binary signal concept (T or M) that's in the middle of two domain concepts with 0-5 general or other concepts separating the domain and signal concept.
(T)[OG]{0,5}(D)[OG]{0,5}(D)	Matches a binary signal concept (T) followed by 0-5 general or other concepts, a domain concept, another 0-5 general or other concepts and a second domain concept.

These patterns were then used to highlight the signal and domain concept(s) in their original context as an additional way of visualizing possible problems in a human readable form.

5.3 Results

For the first approach we generated a new network around the signal concept "difference_ between" to illustrate the usefulness of our approach and specifically the introduction of the new combination nodes. Figure 4 shows this example network with the node for "difference_between" in the middle surrounded by the combination nodes and the referenced domain concept nodes. The red node represents the signal concept this network focuses on while the purple nodes are the new combination. The size of these combination nodes represent their occurrence in all of the texts we analyzed by being larger the more often the combination of signal concept and both domain concepts occurred. Blue nodes show all the domain concepts that were mentioned along with the signal concept "difference_between" in the data we analyzed.

The second approach resulted in a list of comments with co-occurring signal and domain concept(s), containing highlighting for the referenced concepts. An excerpt of this list is shown in table 6, illustrating the usefulness of this approach. It puts our preprocessed text with highlighting next to the original comment to indicate the detection of a signal and domain concept but also shows the original context both as a means of feedback for us showing the validity of our approach but also as a possible source of information for the teacher in a learning analytics context.

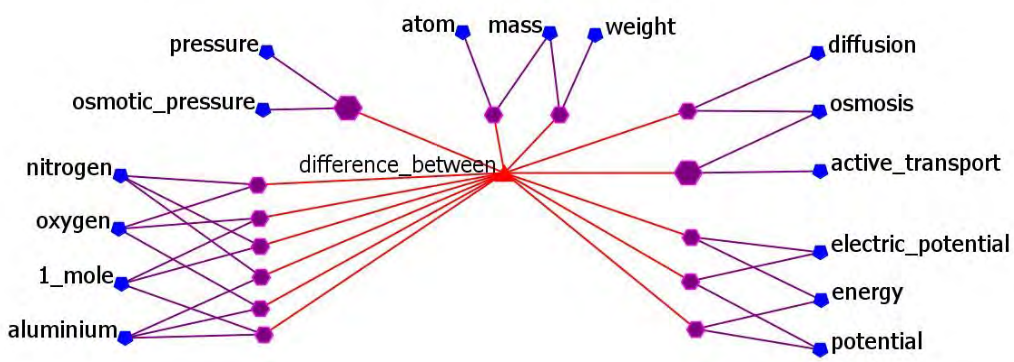

Figure 4. Network including blue combination nodes hovering around "difference_between".

Table 6: Highlighting of signal and domain concepts next to their original context.

I finally understand osmosis. Thanks Khan!! **do_understand osmosis** thanks khan
how I know if the membrane will allow sugar to diffuse or not? plzany body reply. **explanation** i_know if **membrane** be allow sugar diffusion not plzany body reply
KhanAcademy helped me to review a unit on OSMOSIS AND DIFFUSION in my BIOLOGY class! khan_academy **help** review **unit** on osmosis_diffusion in biology class
Still confused about osmotic pressure :/ wasted a bit of time.. still **confusion** about **osmotic_pressure** / waste bite time
What is the difference between osmosis and active transport definition **difference_between osmosis active_transport**

6. Conclusion

In our first study, we aimed to find evidence that NTA is a useful instrument to identify and analyze students' conceptual understanding while learning a specific science topic. Similar to a student's self-created concept map, (semi-)automatically generated text networks provide information about the learners' conceptions belonging to the respective discipline. As theoretically expected, we could corroborate that hubs found in the ensuing concept networks indeed represent ontological root categories, e.g. the concept of "mole" that was the main topic of a role reversal lesson and consequently reaches a high value for total degree. Furthermore, given betweenness centrality values support the theory that nodes high in this measure indicate bridges between different ontological areas. Overall, we conclude that the analysis of a network's structural features provides important information on ontology development. This led us to a second study, in which we investigated the impact of differentiating between various categories.

In the JuxtaLearn context, this type of content analysis can serve as a source of diagnostic information regarding the students' understanding and potential misconceptions. Based on our refinement of the NTA approach, we aim to develop a tool for teachers and educational designers that supports them in optimizing the student's learning processes but also revising educational processes and possibly even curricular decisions.

While investigating general concepts of comments in a discussion around publicly available educational videos, we could identify an ontological subcategory of signal words among frequent general terms high in degree. These words (such as "explanation" or "difference_between") indicate a certain type of concepts, which in conjunction with domain concepts can help identify problems of comprehension. Contrary to commonly used content analysis techniques, which are based on a "bag of words" assumption and hence ignore the word's position in a text (e.g. Sherin, 2012), NTA has proven effective in disclosing conceptual relationships of meaning. By means of extending the NTA by a tailored analysis, we receive indicators for drawing conclusions as to which (pre-)knowledge the students might miss or which concepts are particularly difficult to distinguish (such as "osmosis" and "diffusion").

As a next step on our research agenda, we will particularly look at the progression of concept networks over time. Particularly deviations from "normal" progression following the preferential attachment principle (i.e. disappearing hubs or new clusters of high connectivity) will be analyzed to better understand the evolution of the students' conceptual models.

Acknowledgements

This work was partially funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement 317964 JUXTALEARN.

References

Barabási, A.-L. (2009). Scale-Free Networks: A Decade and Beyond. *Chroma Technology [UNIV DUISBURG-ESSEN]*, 412–413. doi:10.1126/science.1173299

Blei, D. M. (2012). Introduction to Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. doi:10.1145/2133806.2133826

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

Carley, K. M., Pfeffer, J., Reminga, J., Storrick, J., & Columbus, D. (2013). *ORA User's Guide 2013* (Technical Report No. CMU-ISR-13-108) (pp. 1–1280). Pittsburgh: arnegie Mellon University, Institute for Software Research. Retrieved from http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-13-108.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.

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

Engelmann, T., & Hesse, F. W. (2010). How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning*, *5*, 299–319. doi:10.1007/s11412-010-9089-1

Gartner, A., Kohler, M. ., & Riessmann, F. (1971). *Children teach Children: Learning by teaching*. Dallas, TX, U.S.A.: Harper & Row.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, *29*(1), 90–102. doi:10.1016/j.chb.2012.07.020

Heyer, G., Quasthoff, U., & Wittig, T. (2006). *Text Mining: Wissensrohstoff Text*. Bochum: W3L-Verlag, Herdecke.

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.

Hung, J. (2012). Trends of E-learning research from 2000 to 2008: Use of text mining and bibliometrics. *British Journal of Educational Technology*, *43*(1), 5–16. doi:10.1111/j.1467-8535.2010.01144.x

Jacobsen, M. J., & Kapur, M. (2010). Ontologies as Scale Free Networks: Implications for Theories of Conceptual Change. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)* (pp. 193–194).

Julien, H. (2008). Content Analysis. In L. M. Given (Ed.), *Qualitative Research Methods* (pp. 120–121). Thousand Oaks, California: SAGE Publications.

Khan Academy. (2013). A free world-class education for anyone anywhere. *Khan Academy*. Retrieved from https://www.khanacademy.org/about

Meyer, J. H. ., & Land, R. (2003). Threshold concepts and troublesome knowledge: linkages to ways of thinking and practising. In C. Rust (Ed.), *Improving Student Learning - Theory and Practice Ten Years On* (pp. 412–424). Oxford, UK: 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/TheoryUnderly ing ConceptMaps.pdf.

Papert, S., & Harel, I. (1991). *Constructionism*. Norwood, NJ: Ablex Publishing Corporation. Retrieved from http://namodemello.com.br/pdf/tendencias/situatingconstrutivism.pdf

Schreiber, M., & Engelmann, T. (2010). Knowledge and information awareness for initiating transactive memory system processes of computer-supported ad hoc groups. *Computers in Human Behavior*, *26*, 1701–1709. doi:10.1016/j.chb.2010.02.007

Sherin, B. (2012). Using computational methods to discover student science conceptions in interview data. In *LAK '12 Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 188–197). New York: ACM. doi:10.1145/2330601.2330649

Wasserman, S., & and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York (USA): Cambridge University Press.

Wenger, E. (1987). *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.