Aleksander Lempinen

Machine learning based automatic analysis of physics instruction quality

Master's Thesis in Information Technology

September 14, 2020

University of Jyväskylä

Faculty of Information Technology

Author: Aleksander Lempinen

Contact information: aleksander.lempinen@outlook.com

Supervisors: Tommi Kärkkäinen, Daniela Caballero, and Jouni Viiri

Title: Machine learning based automatic analysis of physics instruction quality

Työn nimi: Koneoppimispohjainen automaattinen fysiikan opetuksen laadun analyysi

Project: Master's Thesis

Study line: Educational Technology

Page count: 33+0

Abstract: TODO: Abstract

Keywords: TODO: Keywords

Suomenkielinen tiivistelmä: TODO: Tiivistelmä

Avainsanat: TODO: Avainsanat

Glossary

TODO	TODO: Glossary
ASR	Automatic Speech Recognition
KDD	Knowledge Discovery in Databases
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
NLTK	Natural Language ToolKit
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
CAC	Corrected Arc Curve
QuIP	Quality of Instruction in Physics

List of Figures

List of Tables

Contents

1 Introduction

Automatic speech recognition (ASR) has been one of the topics of interest in fields like computational linguistics and natural language processing, but more recently has expanded into more fields like education. This interest is sparked because of the increasing amount of audio speech data gathered in research activities such as observation or interviews. While speech-to-text ASR has been available for some time, further analysis of the text data in has remained a manual process especially in the domain of education. An important part of physics education is teaching students how to think like an expert and have deeper understanding of the physics concepts with monitoring and guidance of the instructor (Wieman and Perkins 2007). Good physics instruction depends on the talk between the teacher and the students, which is called *teacher talk* (Scott and Ametller 2007). Unfortunately teacher talk is not given the attention it deserves during teacher education (Crespo 2002; Lehesvuori 2013).

Speech as data is sequential and messy in nature and requires cleaning and preprocessing before it can be used. In the case of teacher talk, it is traditionally analysed manually by researchers, which is a subjective and laborious process that is difficult to reproduce. Lack of analysis methods for teacher talk have been identified and new methods to analyse and visualize teacher talk during a lesson have been developed over time (Viiri and Saari 2006; Lehesvuori et al. 2013). However, these methods rely on researchers first manually transcribing and coding the data, which is still laborious and difficult to reproduce. This is impractical for both research and for use in the field to give teachers feedback during their teacher education. Speech-to-text ASR is a well researched topic with significant advances especially in languages without a lot of speakers such as Finnish (Kurimo et al. 2017). Further processing of the text data is domain and application specific and not well researched in the case of teacher talk in science education and is the topic of this thesis.

The problem is approached through a *Knowledge Discovery in Databases (KDD)* process, which is an iterative process developed to extract valid, novel, potentially useful and understandable patterns from existing databases when manual data analysis is slow, expensive, highly subjective or otherwise impractical (Fayyad, Piatetsky-Shapiro, and Smyth 1996c).

KDD is a formal process that in addition to a *data mining* step includes data preparation and interpretation and evaluation steps (Fayyad, Piatetsky-Shapiro, and Smyth 1996c). This type of data analysis is exploratory in nature without assumptions and is used for automatic hypothesis generation, with emphasis on evaluation using quantitative measures (Fayyad, Piatetsky-Shapiro, and Smyth 1996c).

The aim of this thesis is to automatically discover novel and potentially useful patterns from transcripts of teacher talk during physics lessons using data mining methods.

The outcomes of this thesis are two contributions in analysing teacher talk:

- Automatic visualization of teacher talk using concept networks
- motif discovery

2 Physics instruction quality

TODO introduction to the chapter

2.1 Teacher talk

The quality of physics instruction has been at the centre of discussion, where traditional teaching methods have been criticized as inefficient at creating experts capable of thinking like physicists (Wieman and Perkins 2007). The interaction between the teacher and the students is called *teacher talk* and is a big part of all physics lessons, which is often taken for granted (Scott and Ametller 2007). Teacher talk is often not sufficiently addressed during teacher education and could be explained by the scarcity of available methods (Lehesvuori 2013; Viiri and Saari 2006; Crespo 2002).

2.2 Qualitative analysis of teacher talk

Qualitative observation, where a mentor teacher makes observations during the lesson and gives feedback to the student teacher after the lesson, is an easy and a natural way to analyse teacher talk. According to Viiri and Saari (2006) student teachers have issues with remembering what happened during the lesson and both self-reflection and teacher tutor feedback are based on memory and are unstructured. Therefore they developed a method to analyse teacher talk by visualizing talk types such as "teacher presentation", "authoritative discussion", "dialogic discussion", "peer discussion" and "other" of the lesson. This was achieved by first videotaping the lesson, manually transcribing it and then manually coding each time window into one of the teacher talk types. Viiri and Saari (2006) point out that ”Discourse analysis necessarily proceeds on the basis of the investigator’s interpretations of what was said.” This newly developed method is qualitative and subjective in nature by necessity.

This kind of qualitative and subjective analysis of teacher talk from classroom videos, usually by first transcribing or coding the talk based on some kind of a theoretical framework is not unique. For example Scott and Ametller (2007) and Scott et al. (2011) rely on small case

studies and manual analysis of teacher talk from classroom video data either directly or from transcripts. The details of what they are looking for in the teacher talk is different in each case and dependent on the chosen theoretical framework and the research questions, but overall the process is similar.

Figure 1 represents a typical pipeline for analysis of teacher talk. First a meaningful representation is coded directly from transcriptions and classroom video. The representations are then visualized and interpreted or interpreted directly. Transcripts are often partial, because obtaining accurate transcriptions requires a lot of work and researchers will prefer to only transcribe some of the video and perform coding directly on the video. This type of coding is usually done by a single researcher. For example Jokiranta (2014) noted in her Master's thesis that the codings of the teacher talk data in her thesis were driven by research questions and literature and the analysis is based on the author's interpretation.

Qualitative analysis of teacher talk is not necessarily a weakness. Lehesvuori (2013) used a mix of qualitative and quantitative methods in his PhD dissertation and notes that qualitative methods of analysing teacher talk have allowed for more flexibility in contrast with quantitative methods, which were limited by the scarcity of dialogic interactions during the lessons. He identifies a weakness with quantitative methods of analysing teacher talk, which typically do not take into account the temporal aspect of teacher talk continuously changing during the lesson. Lehesvuori (2013) also developed a method to visualize the teacher talk from classroom video of the lesson similar to Viiri and Saari (2006) under a different theoretical framework.

2.3 Quantitative analysis of teacher talk

A more quantitative approach was used by Helaakoski and Viiri (Helaakoski and Viiri 2014) to analyse the content structure of teacher talk. They used a relatively large dataset of classroom video consisting of 45 German, 28 Swiss and 25 Finnish lessons about "Relation between electrical energy and power." The videos were manually transcribed and from the videos and transcripts links between concept categories were identified and represented as a connectivity matrix, which could be visualized as a network of concepts. From the same

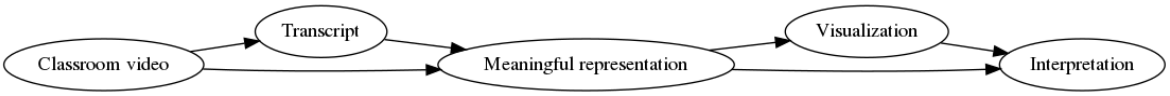

Figure 1. Teacher talk analysis has a common manual pipeline. Classroom video is transcribed, meaningfully coded and then visualized and interpreted.

connectivity other metrics and measures were computed. Helaakoski and Viiri (2014) found that ”More specifically, the frequencies of physics concepts and connections between them correlated significantly with learning gains.” This result is in line with previous research, which stresses the importance of paying attention to teacher talk. (Viiri and Saari 2006; Scott and Ametller 2007; Scott, Mortimer, and Ametller 2011)

In summary, teacher talk is an important component of any lesson and it has not been sufficiently researched, because of a lack of good research methods. Specifically analysis of the temporal aspect of teacher talk during the lesson is interesting, but has few methods to analyse it. Improved methods to analyse teacher talk could lead to better feedback to student teachers and more opportunities for researching the interactions of the teacher during a lesson.

3 Teacher talk as data

TODO: Introduction to the chapter

3.1 Natural language processing

Natural language processing as a field deals with applications of human language data, which is different from computational linguistics which focuses on studying human languages, but there is overlap and sharing of methods between the fields.

3.1.1 Automatic speech recognition

Automatic speech recognition (ASR) can be thought of as a subset of natural language processing. In fact, most ASR systems are speech-to-text, where any further processing is done with the text data. Speech might not always have clearly defined words and sentences, which shifts the focus from words to sounds called *phonemes* and from sentences to word sequences called *utterances*. Modelling conversational speech is especially difficult (Kurimo et al. 2017).

ASR is a classification task, where the goal is to predict what was said from the audio signal of speech. Early ASR systems had an acoustic model which detected phonemes to recognize numbers, some vowels and consonants for a single speaker (Juang and Rabiner 2005). Improvements in the acoustic model allowed for introduction of speaker-independent ASR (Benzeghiba et al. 2007; Juang and Rabiner 2005). The later addition of a language model based on statistical grammar and syntax helped more accurately predict the correct word based on what words previously appeared in the utterance (Juang and Rabiner 2005). Modern ASR systems utilize the fact that utterances are sequences of words and words are sequences of phonemes (Bengio and Heigold 2014).

Most commonly ASR is based on sequence models such as *Hidden Markov Models (HMM)*, but deep learning approaches using *Recurrent Neural Networks (RNN)* and *Long Short-Term Memory (LSTM)* networks are gaining popularity for both acoustic models and language

models or even end to end text-to-speech models (Bengio and Heigold 2014; Enarvi et al. 2017). Deep learning approach is more data-driven and relies on fewer assumptions, but instead requires more data for training (Bengio and Heigold 2014). This might be impractical if training data is limited, which is often the case with languages without a lot of speakers such as Finnish. Depending on the architecture, an ASR system might be capable of either transcription, keyword spotting or both (Juang and Rabiner 2005; Enarvi et al. 2017).

ASR is a difficult machine learning task because of a large search space, large vocabulary, undetermined length of word sequences and problems related to aligning speech signal to the text (Enarvi et al. 2017). Speech is highly variable even with a single speaker due to noise, but different pronunciations and accents mean that the audio signal will be different despite the same words being spoken (Juang and Rabiner 2005). Accents, dialects, emotional state, gender and casual speech slurring in spontaneous speech bring a lot of variation which makes conversational speech especially difficult compared to standard pronunciations and vocabulary (Benzeghiba et al. 2007; Juang and Rabiner 2005). Speaker-dependent systems are typically more accurate than speaker-independent systems (Benzeghiba et al. 2007; Enarvi et al. 2017).

3.1.2 Text processing

TODO: NLP overview (Silfverberg et al. 2016; Kanerva et al. 2018)

TODO: Definitions of concepts such as term, keyword, document, corpus and text

3.1.3 Finnish language

TODO: Conversational vs official Finnish

TODO Finnish language is particularly difficult for ASR because words are formed by concatenating smaller (Enarvi 2018)

4 Methods

According to Fayyaad et. al (1996a) traditional methods of manual analysis and interpretation rely on the person performing the analysis to be very familiar with the data and is a slow, expensive and highly subjective process. Even in the 90's they have noticed, that "manual data analysis is becoming completely impractical in many domains". Because of the fast pace of accumulation of new data, they identified an urgent need for computational theories and tools to extract knowledge from data in at least a partially automatic way.

Quality of Instruction in Physics (QuIP) project (Fischer et al. 2014) used this type of manual data analysis process to extract patterns of instruction and find surface and deep structures in instructional quality in physics. The main research questions in the QuIP study were the following:

1. What are the typical patterns of physics instruction?
2. Under what conditions are these patterns successful with respect to student learning, interest and motivation?

Answering the first research question required finding patterns using manual data analysis, because of unavailability of automatic methods in teacher talk research. This thesis uses the audio data collected from teachers during the QuIP project to automatically extract patterns physics instruction using the Knowledge Discovery in Databases (KDD) process.

4.1 Dataset

The dataset consists of audio recordings of Finnish physics teachers in comprehensive schools collected during the Quality of Physics Instruction study (Fischer et al. 2014; Helaakoski and

Start	**End**	**Text**
10.0	15.0	hehkulamppua kakstoista tunteesta kirjottaas taas että miten tuo laske
15.0	20.0	virtapiirejä virtapiirejä rakettien kans vaikka kuinka paljon tähän

Table 1. An example of 5 second segments obtained from an ASR system

Viiri 2014). Helaakoski and Viiri (2014) used microphones attached to 25 physics teachers to record lessons on the topic of ”relation between electrical energy and power”. Of the 25 teachers, 22 taught their topic during two 45 min lessons and 3 teachers taught their topic during a 90 minute double lesson.

In addition to audio recordings from the teachers, the students were tested on the knowledge of the topic before and after the lesson and a list of physics keywords was obtained from a ”FyKe 7-9 Fysiikka”(Kangaskorte et al. 2016) physics textbook glossary. The audio data was then automatically transcribed using AaltoASR into 5 second segments (Hirsimaki, Pylkkonen, and Kurimo 2009) as shown in table 1.

TODO Dataset quality evaluation criteria

4.2 KDD process

To extract knowledge in an automatic way a new field called Knowledge Discovery in Databases (KDD) was introduced, that combines the fields of expert systems, machine learning, intelligent databases, knowledge acquisition, case-based reasoning and statistics (Piatetsky-Shapiro 1990). Originally KDD referred to the interactive and iterative process of extracting useful knowledge from data with emphasis on that knowledge is the end product of a data-driven discovery (Fayyad, Piatetsky-Shapiro, and Smyth 1996b). Data mining was a particular step in the process (Fayyad, Piatetsky-Shapiro, and Smyth 1996b), but later data mining and KDD became synonymous (Piatetsky-Shapiro 2000). A formal definition provided by Fayaad et. al (1996b) is as follows:

> Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

TODO How is KDD different from regular data analysis?

TODO Importance of formulating the problem

Found patterns should be novel, understandable, which are more subjective measures and valid on new data, which has more objective measures to evaluate them (Fayyad, Piatetsky-

Word	Snowball stem	Voikko lemma	Translation
virta	vir	Virta	current
virtaa	virt	virrata	current
virtapiiri	virtapiir	virtapiiri	electric circuit
virtapiirejä	virtapiir	virtapiiri	electric circuits

Table 2. An example of a stemming and lemmatization error with variations of the word "current"

Shapiro, and Smyth 1996b). This evaluation is important, because knowledge discovery is based in statistics and patterns can be found even in random data (Fayyad, Piatetsky-Shapiro, and Smyth 1996a).

The KDD process can be broken down into nine steps with iterations and loops possible between any two steps (Fayyad, Piatetsky-Shapiro, and Smyth 1996c, 1996b, 1996a):

1. Understanding the domain
2. Creating target dataset
3. Data cleaning and preprocessing
4. Data reduction and projection
5. Matching goals to data mining method
6. Exploratory data analysis
7. Data mining
8. Interpreting mined patterns
9. Acting on discovered knowledge

TODO Steps that are the focus of this thesis: preprocessing, data reduction & projection

4.3 Text Normalization

First we apply text normalization techniques to the documents. First the documents are tokenized split by split into sequences of words using the NLTK tokenizer (Bird, Klein, and Loper 2009). Because of the agglutinative nature of Finnish, we then use the NLTK Snowball stemmer (Bird, Klein, and Loper 2009) and the Voikko morphological analyser

(Pitkänen 2019) for stemming and lemmatization. We then process the physics keyword list in the same way.

Stemming relies on an algorithm to remove suffixes and prefixes to get the same stem from slightly different words that have roughly same meaning such as *present* and *presented*. On the other hand, lemmatization relies on a dictionary of known variations of the word and can handle cases problematic for a stemmer, when the spelling of the word is different between variations such as *drink* and *drank*. Stemming is a faster and simpler approach compared to lemmatization (Flores and Moreira 2016). Neither stemming or lemmatization works in cases, when a variation of the word might have different options for the base word, an example of which is shown in table 2.

According to Flores and Moreira (2016), stemming accuracy did not improve information retrieval in English, but did improve in languages such as Portuguese, Spanish and French. Stemmers can be evaluated intrinsically, which is evaluating only the stemming system or extrinsically, which is evaluating the effect of the stemmer on the entire application (Flores and Moreira 2016).

During exploratory analysis of the dataset intrinsic evaluation of the stemmer using the multiple edit distance metrics provided by NLTK (Bird, Klein, and Loper 2009) found that stemming does not work well with Finnish, because of the large amount of word variations. Despite this we would like to evaluate the effects of both the stemming and the lemmatization approach on the entire system using extrinsic evaluation. In other words, we compare the end results of the entire pipeline with stemming, lemmatization and doing nothing.

4.4 Bag-of-words

A bag-of-words model is a simple way to represent a document as a feature, where the position of the words is ignored (Jurafsky and Martin 2019). The basic idea is that for each token in the document we count the number of times the token occurred and represent it as a feature. For example two documents consisting of the sentences ”the, pen, is, in, the, bag” and ”a, pen, is, mightier, than, a, sword” would be represented like in table 3. Other variations include using *term frequency (TF)*, where the token count is divided by the

Document	**a**	**bag**	**in**	**is**	**mightier**	**than**	**the**	**pen**	**sword**
Document 1	0	1	1	1	0	0	2	1	0
Document 2	1	0	0	1	1	1	0	1	1

Table 3. Example of bag of words with counts

total amount of tokens in the document to reduce the impact of differences in document length or *term frequency - inverse document frequency (TF-IDF)*, where the term frequency is multiplied by inverse document frequency to reduce the impact of uninteresting words that commonly occur in many documents (Leskovec, Rajaraman, and Ullman 2020).

Bag of words models create very large feature vectors with as many features as there are unique tokens in the corpus. With a large enough vocabulary the data is incredibly sparse with most of the tokens occurring zero times in a given document. This sparsity of the data can create problems further in the analysis pipeline due to the high dimensionality, especially if there are few documents (Verleysen and François 2005). To decrease the sparsity, we can combine the features into bins and cutting the long tail (Leskovec, Rajaraman, and Ullman 2020; Jurafsky and Martin 2019). A common technique is to combine the very rare tokens into an ”other” feature and to remove very common tokens called *stop words* that do not contain a lot of information such as ”is, a, the, and, where” (Jurafsky and Martin 2019).

During exploratory analysis we quickly noticed that it is necessary to reduce the amount of features. To do this we focus on the keyword list and combine tokens outside of the keyword list into an *other* feature. This leaves us with a few hundred features instead of thousands of features. To reduce the amount of features even more, we bin the keywords into the following categories with the help of a physics expert: concepts, objects, units and meta keywords. Since some teachers are expected to talk more than others, we apply the TF technique by dividing the counts of keywords spoken by the total amount of words spoken by the teacher.

To evaluate this type of feature engineering, we use the same approach as in the extrinsic evaluation of the stemming and lemmatization during text normalization.

4.5 Token co-occurrence

TODO Describe token co-occurrence (Jurafsky and Martin 2019; Bullinaria and Levy 2007) TODO Describe weighted co-occurrence TODO Model evaluation criteria

4.6 Concept network analysis

Expanding upon the idea of visualizing concept structure using conceptual networks by Helaakoski and Viiri (2014), we automatically generate weighted networks using the physics keyword list. In addition to obtaining a visualization of the content structure of the teacher talk during the lesson, we compute network measures to quantify the relationships between the concepts.

We first compute a physics concept co-occurrence matrix for each text normalized lesson using an overlapping window over the splits produced by the ASR system. Since a non-overlapping window would not count two concepts occurring together but in different windows as co-occurring, we count the co-occurrences within the first half of the window and then add the co-occurrences between the two halves. This is done to avoid counting the same co-occurrence within a window twice due to the overlapping window.

We then represent a pair of co-occurring physics concepts as nodes with the amount of co-occurrences as the weight for that edge. Using NetworkX (Hagberg, Schult, and Swart 2008) we can visualize the lessons concept structure as a network of concepts as shown in figure

TODO Model evaluating criteria

4.7 Motif and discord discovery

There are two main approaches to representing the temporal aspect of a lesson. The first approach would be to engineer more features, for example instead of computing a mean for the entire lesson we would split the lesson into smaller pieces called *bins* and compute a separate mean for each bin. The second approach would be to have a sequence of examples, each representing a piece of the lesson. Such a sequence consisting of real umbers is called a time

series (Yeh, Kavantzas, and Keogh 2017). Data mining and knowledge discovery of time series data involves techniques such as *motif* discovery, *discord* discovery, and *segmentation* (Yeh 2018).

A motif is a pattern, where a sequence is repeating such as a heartbeat or an annual spike of sales before Christmas. On the other hand, a discord is a sequence that is different from the rest of the time series such as anomalies or black-swan events. More precisely, a motif is a pair of local subsequences that are very similar compared to the rest of the time series and a discord is a local subsequence that is very dissimilar compared to the rest of the time series (Yeh 2018). Semantic segmentation splits the time series into meaningful segments, where the patterns change. For example splitting a heartbeat time series into a "rest" and "run" segments.

Motif discovery, discord discovery and semantic segmentation can be done using an unsupervised technique called Matrix Profile (Yeh 2018). The basic idea of a Matrix Profile is to compute the smallest Euclidean distance of the local sequence to some other sequence elsewhere in the time series for each point in the time series. The result is that with only a parameter of window width we can obtain motifs that have the same low distance value, discords that are the peaks and have a large distance value and semantic segmentation that have few motifs spanning between the segments. There are three distinct ways to apply the Matrix Profile to analysing teacher talk.

The first way is to use the keyword frequency time series to compute a matrix profile for each lesson and use motif discovery to find patterns of similar local sequences within the same lesson. A *corrected arc curve (CAC)* gives us the amount of motif pairs that cross each point. The lowest point of the CAC will have the fewest motifs spanning across it and it can be used to segment the lesson into two parts that have different patterns within them.

The second way is to append all keyword frequency time series into a single time series, compute the matrix profile and use motif discovery to find patterns of similar local sequences that can be between lessons or between different teachers.

The third way is to have each lesson as a separate dimension and compute a multidimensional matrix profile and use motif discovery to find patterns that occur in the same point of the

lesson across two or more lessons.

After the motif and discord discovery and semantic segmentation is completed, we visualize the matrix profile, CAC and the found patterns and interpret the local sequences. This interpretation phase is a manual process that relies on domain expertise and is necessary when using unsupervised techniques.

TODO Model evaluation criteria

4.8 Cluster analysis

TODO Unsupervised technique, based on a distance measure, types of distance measures, curse of dimensionality

TODO Hierarchical clustering, AgglomerativeCLustering, Dendogram, linkage method

TODO Model evaluation criteria: CLuster validation indices, interpretation of the clusters

4.9 Predictive modelling

According to Shmueli (2010) predictive modelling is defined as ”the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations”. This is in contrast to explanatory modelling, where causality provided by the theory is tested using statistical methods (Shmueli 2010). It is important to point out, that explanatory modelling relies on *construct operationalization*, that is tying theoretical constructs to observable measurements, but the predictive modelling does not (Shmueli 2010). In other words, explanatory modelling aims to extract information about how nature associates inputs to outputs, while predictive modelling aims to use inputs to predict outputs (Breiman 2001). TODO Supervised technique, example and label

TODO Decision tree

TODO Model evaluation criteria: validation, holdout set, k-fold, comparison to naive algorithms

4.10 Evaluation and interpretation

5 Results

5.1 Concept network analysis

5.2 Matrix profile

TODO Example matrix profile and the found motifs, discords and segmentation

TODO Interpretation of the matrix profile motifs, discords and segmentation

TODO Instability related to parameters and seemingly spurious patterns

5.3 Cluster analysis

TODO Dendogram of the clusters, results of cluster validation indices

TODO Interpretation of clusters TODO

TODO Bad quality of the clusters, low cluster validation score

5.4 Predictive modelling

TODO Chosen label and performance metric, results of the trained models using k-fold TODO

TODO Example of a decision tree

Performance of the decision tree model compared to a classifier that picks a class at random on previously unseen data using 10-fold validation can be seen in table 4. The performance

Method	**Train accuracy**	**Test accuracy**
Decision tree	NA	NA
Random classifier	NA	NA

Table 4. Performance comparison of different predictive methods

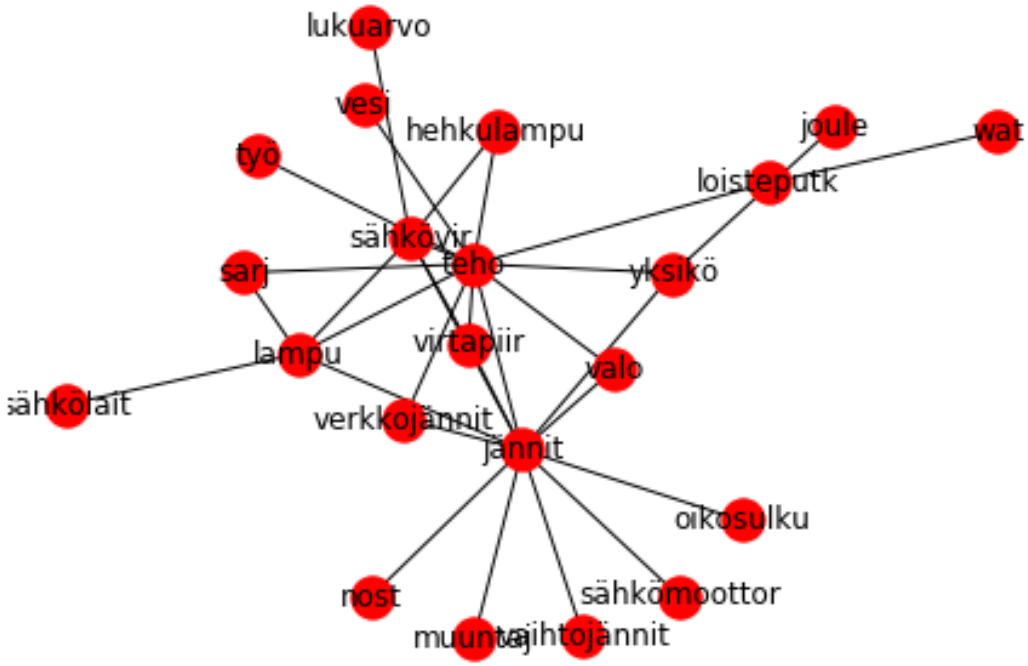

Figure 2. Concept network obtained from teacher talk during a Finnish physics lesson on the topic of

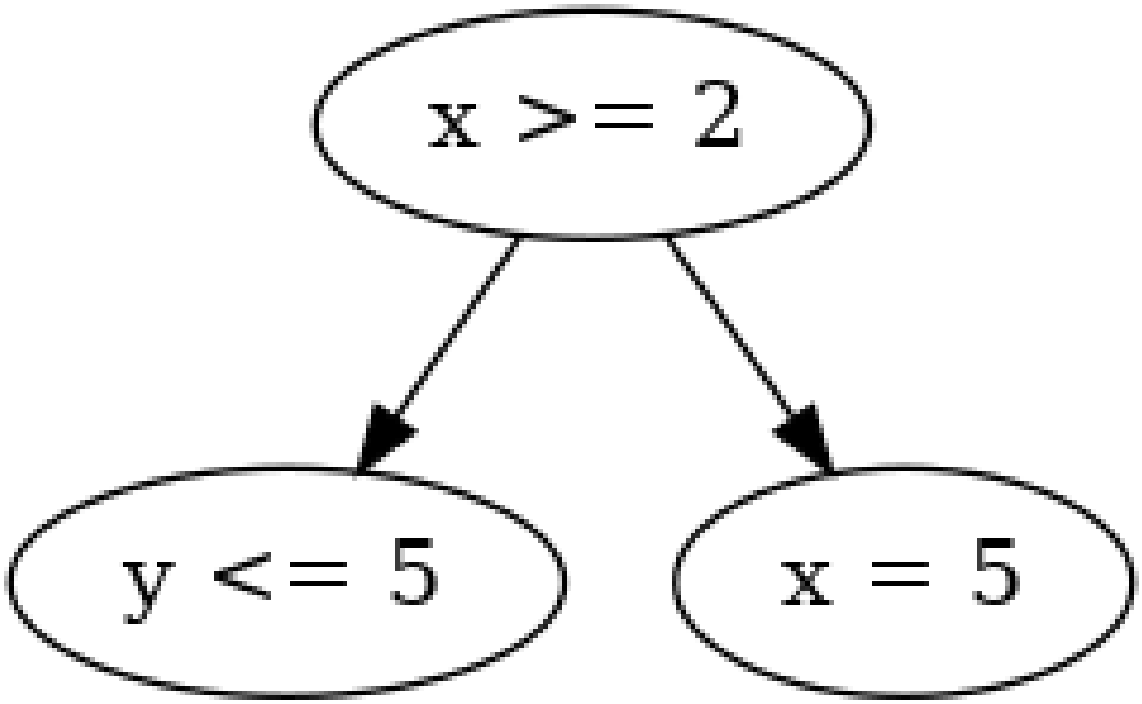

Figure 3. The decision tree obtained when predicting whether it is the beginning, the middle or the end of the lesson

is very similar to a random classifier, which means that the model could not extract any useful patterns from the data.

6 Related work

Caballero et al. (2017) applied ASR and NLP techniques to analyse teacher talk in Spanish and Finnish in a pilot study to determine the feasibility of the approach and maturity of modern ASR systems. Two teachers were equipped with microphones and concept networks were created and visualized by counting which physics concept keywords occurred together in the teacher's talk. We expanded upon this idea by using a larger dataset with more teachers, using alternative NLP methods such as lemmatization in addition to stemming, adding more representations of the data and including taking the temporal aspect of the lesson into account in the analysis and visualizations.

Fan et al. (2015) used NLP techniques to improve interactions between the student and the instructor. They developed a pilot system to collect self-reflections from the students, summarize them using NLP and then show the summaries to the instructors through a graphical interface. We believe that a simple to use interface to show summaries is necessary for teachers and developed visualizations so that teachers can see the content structure of the entire lesson at a glance.

TODO Find more related work

7 Discussion

TODO partial success of token co-occurrence: network analysis TODO Failure of bag of words: matrix profiles, clustering and decision trees

TODO interpretation of found patterns

TODO Acting on discovered knowledge

TODO Limitations

TODO Ethics and privacy In addition to the ethics of data collection in a classroom setting and potential misuse, automatic processing of teacher talk could be used for teacher assessment that is not based in pedagogical theory. The methods described in this thesis are based on potentially useful patterns in the data and further confirmatory research is required to determine whether the patterns are indeed useful and how they align with pedagogical theory.

8 Conclusions

TODO Further research

TODO Acknowledgements Research was done in collaboration with the Department of Teacher Education at University of Jyväskylä and Centro de Investigación Avanzada en Educación (CIAE) at University of Chile.

Bibliography

Bengio, Samy, and Georg Heigold. 2014. ”Word Embeddings for Speech Recognition”. In *Proceedings of the 15th Conference of the International Speech Communication Association, Interspeech.*

Benzeghiba, M., R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, et al. 2007. ”Automatic Speech Recognition and Speech Variability: A Review” [**inlang**en]. *Speech Communication,* Intrinsic Speech Variations, 49, number 10 (): 763–786. ISSN: 0167-6393. doi:10.1016/j.specom.2007.02.006.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* [**inlang**en]. "O'Reilly Media, Inc." ISBN: 978-0-596-55571-9.

Breiman, Leo. 2001. ”Statistical Modeling: The Two Cultures” [**inlang**en]. *Statistical Science* 16 (3): 199–215.

Bullinaria, John A., and Joseph P. Levy. 2007. ”Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study” [**inlang**en]. *Behavior Research Methods* 39, number 3 (): 510–526. ISSN: 1554-3528. doi:10.3758/BF03193020.

Caballero, Daniela, Roberto Araya, Hanna Kronholm, Jouni Viiri, André Mansikkaniemi, Sami Lehesvuori, Tuomas Virtanen, and Mikko Kurimo. 2017. ”ASR in Classroom Today: Automatic Visualization of Conceptual Network in Science Classrooms” [**inlang**en]. In *Data Driven Approaches in Digital Education,* edited by Élise Lavoué, Hendrik Drachsler, Katrien Verbert, Julien Broisin, and Mar Pérez-Sanagustín, 541–544. Lecture Notes in Computer Science. Cham: Springer International Publishing. ISBN: 978-3-319-66610-5. doi:10.1007/978-3-319-66610-5_58.

Crespo, Sandra. 2002. ”Praising and Correcting: Prospective Teachers Investigate Their Teacherly Talk” [**inlang**en]. *Teaching and Teacher Education* 18, number 6 (): 739–758. ISSN: 0742-051X. doi:10.1016/S0742-051X(02)00031-8.

Enarvi, Seppo. 2018. *Modeling Conversational Finnish for Automatic Speech Recognition* [**inlang**en]. Aalto University. ISBN: 978-952-60-7908-0.

Enarvi, Seppo, Peter Smit, Sami Virpioja, and Mikko Kurimo. 2017. "Automatic Speech Recognition with Very Large Conversational Finnish and Estonian Vocabularies" [**inlang**en] (). doi:10.1109/TASLP.2017.2743344.

Fan, Xiangmin, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. 2015. "CourseMIRROR: Enhancing Large Classroom Instructor-Student Interactions via Mobile Interfaces and Natural Language Processing". In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems,* 1473–1478. CHI EA '15. Seoul, Republic of Korea: Association for Computing Machinery. ISBN: 978-1-4503-3146-3. doi:10.1145/2702613.2732853.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996a. "From Data Mining to Knowledge Discovery in Databases" [**inlang**en]. *AI Magazine* 17, number 3 (): 37–37. ISSN: 2371-9621. doi:10.1609/aimag.v17i3.1230.

———. 1996b. "Knowledge Discovery and Data Mining: Towards a Unifying Framework" [**inlang**en]. In *KDD-96 Proceedings,* 7.

———. 1996c. "The KDD Process for Extracting Useful Knowledge from Volumes of Data" [**inlang**en]. *Communications of the ACM* 39, number 11 (): 27–34. ISSN: 00010782. doi:10.1145/240455.240464.

Fischer, Hans E, Peter Labudde, Knut Neumann, and Jouni Viiri. 2014. *Quality of Instruction in Physics: Comparing Finland, Germany and Switzerland* [**inlang**English]. Münster: Waxmann Verlag. ISBN: 978-3-8309-3055-6.

Flores, Felipe N., and Viviane P. Moreira. 2016. "Assessing the Impact of Stemming Accuracy on Information Retrieval – A Multilingual Perspective" [**inlang**en]. *Information Processing & Management* 52, number 5 (): 840–854. ISSN: 0306-4573. doi:10.1016/j.ipm.2016.03.004.

Hagberg, Aric A, Daniel A Schult, and Pieter J Swart. 2008. "Exploring Network Structure, Dynamics, and Function Using NetworkX" [**inlang**en]: 5.

Helaakoski, Jussi, and Jouni Viiri. 2014. ”Content and Content Structure of Physics Lessons and Students' Learning Gains: Comparing Finland, Germany and Switzerland” [**inlang**English]. In *Quality of Instruction in Physics: Comparing Finland, Germany and Switzerland,* 93–110. Münster: Waxmann Verlag. ISBN: 978-3-8309-3055-6.

Hirsimaki, Teemu, Janne Pylkkonen, and Mikko Kurimo. 2009. ”Importance of High-Order N-Gram Models in Morph-Based Speech Recognition”. *IEEE Transactions on Audio, Speech, and Language Processing* 17, number 4 (): 724–732. ISSN: 1558-7924. doi:10.1109/TASL.2008.2012323.

Jokiranta, Kaisa. 2014. ”The Nature of Teacher Discourse during Practical Work in Lower Secondary Physics Education” [**inlang**eng].

Juang, B H, and Lawrence R Rabiner. 2005. ”Automatic Speech Recognition – A Brief History of the Technology Development” [**inlang**en]. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1:24.

Jurafsky, Daniel, and James H. Martin. 2019. *Speech and Language Processing* [**inlang**English]. 3rd ed. draft.

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. ”Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task”. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,* 133–142. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/K18-2013.

Kangaskorte, Anne, Jari Lavonen, Outi Pikkarainen, Heikki Saari, Jarmo Sirviö, Kirsi-Maria Vakkilainen, and Jouni Viiri. 2016. *FyKe 7 - 9 Fysiikka (OPS 2016)* [**inlang**fi]. Fourteenth. Sanoma Pro. ISBN: 978-952-63-1360-3.

Kurimo, Mikko, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi, and Tanel Alumäe. 2017. ”Modeling Under-Resourced Languages for Speech Recognition” [**inlang**en]. *Language Resources and Evaluation* 51, number 4 (): 961–987. ISSN: 1574-0218. doi:10.1007/s10579-016-9336-9.

Lehesvuori, Sami. 2013. ”Towards Dialogic Teaching in Science : Challenging Classroom Realities through Teacher Education” [**inlang**eng]. *Jyväskylä studies in education, psychology and social research,* number 465.

Lehesvuori, Sami, Jouni Viiri, Helena Rasku-Puttonen, Josephine Moate, and Jussi Helaakoski. 2013. ”Visualizing Communication Structures in Science Classrooms: Tracing Cumulativity in Teacher-Led Whole Class Discussions” [**inlang**en]. *Journal of Research in Science Teaching* 50 (8): 912–939. ISSN: 1098-2736. doi:10.1002/tea.21100.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of Massive Data Sets.* Cambridge university press.

Piatetsky-Shapiro, Gregory. 1990. ”Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop” [**inlang**en]. *AI Magazine* 11, number 4 (): 68–68. ISSN: 2371-9621. doi:10.1609/aimag.v11i4.873.

———. 2000. ”Knowledge Discovery in Databases: 10 Years After”. *ACM SIGKDD Explorations Newsletter* 1, number 2 (): 59–61. ISSN: 1931-0145. doi:10.1145/846183.846197.

Pitkänen, Harri. 2019. *Voikko – Free Linguistic Software for Finnish.* https://voikko.puimula.org/.

Scott, Phil, and Jaume Ametller. 2007. ”Teaching Science in a Meaningful Way: Striking a Balance between ’opening up’ and ’Closing down’ Classroom Talk”.

Scott, Phil, Eduardo Mortimer, and Jaume Ametller. 2011. ”Pedagogical Link-making: A Fundamental Aspect of Teaching and Learning Scientific Conceptual Knowledge”. *Studies in Science Education* 47, number 1 (): 3–36. ISSN: 0305-7267. doi:10.1080/03057267.2011.549619.

Shmueli, Galit. 2010. ”To Explain or to Predict?” [**inlang**en]. *Statistical Science* 25, number 3 (): 289–310. ISSN: 0883-4237. doi:10.1214/10-STS330.

Silfverberg, Miikka, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2016. ”FinnPos: An Open-Source Morphological Tagging and Lemmatization Toolkit for Finnish” [**inlang**en]. *Language Resources and Evaluation* 50, number 4 (): 863–878. ISSN: 1574-0218. doi:10.1007/s10579-015-9326-3.

Verleysen, Michel, and Damien François. 2005. "The Curse of Dimensionality in Data Mining and Time Series Prediction" [**inlang**en]. In *Computational Intelligence and Bioinspired Systems,* edited by Joan Cabestany, Alberto Prieto, and Francisco Sandoval, 758–770. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. ISBN: 978-3-540-32106-4. doi:10.1007/11494669_93.

Viiri, Jouni, and Heikki Saari. 2006. "Teacher Talk Patterns in Science Lessons: Use in Teacher Education" [**inlang**en]. *Journal of Science Teacher Education* 17, number 4 (): 347–365. ISSN: 1573-1847. doi:10.1007/s10972-006-9028-1.

Wieman, Carl, and Katherine Perkins. 2007. "Transforming Physics Education" [**inlang**en]. *Physics Today* 58, number 11 (): 36. ISSN: 0031-9228. doi:10.1063/1.2155756.

Yeh, Chin-Chia Michael. 2018. "Towards a Near Universal Time Series Data Mining Tool: Introducing the Matrix Profile" [**inlang**en]. *arXiv:1811.03064 [cs, stat]* (). arXiv: 1811.03064 [cs, stat].

Yeh, Chin-Chia Michael, Nickolas Kavantzas, and Eamonn Keogh. 2017. "Matrix Profile VI: Meaningful Multidimensional Motif Discovery" [**inlang**en]. In *2017 IEEE International Conference on Data Mining (ICDM),* 565–574. New Orleans, LA: IEEE. ISBN: 978-1-5386-3835-4. doi:10.1109/ICDM.2017.66.