Multimodal Data Analysis for Online Broadcasting

Thesis Project 2. Structured Prediction for Semantic Analysis and Knowledge Representation

1 Main goals

Language is increasingly analysed in a cross-media context, where the derived information is part of a greater whole. Consider a news publicist who is compiling a feature video on Elon Musk. She search her agency's archive for "Video footage in which Musk talks excitedly about his rocket's succesful reentry into earth atmosphere". Large agencies such as Reuters produce hundreds of edited videos per day, and the raw footage is magnitudes greater. It is therefore not efficient for our publicist to manually search the archive. Instead, she wants to query the archive using multimodal data analysis, which in this particular case requires a coordinated application of speech recognition, face detection, emotion recognition, and topic identification. The purpose of the MICO platform [Aic+15], is to let her do exactly that.

MICO, short for Media in Context, is a multimodal analysis platform developed in a joint collaboration between Fraunhofer IDMT, University of Passau, Oxford University, Salzburg Research and Umeå University. With MICO, the publicist can analyse, search, and query the archive, and in addition generate recommendations of related content for a given media item. The analysis is done by a battery of autonomous metadata extractors which are orchestrated by a central information broker. The metadata resulting from the analysis is stored in the semantic web standard format RDF (Resource Description Format). The publicist can interact with the system through her browser, and save the results she considers most useful to her local disk.

The MICO project was successfully completed in November 2016. In their concluding technical report, the reviewers emphasized that the project had fully achieved its objectives and technical goals and had even exceeded expectations [TG16]. While it is true that we solved the problems staked out in the beginning of the project, we also discovered new challenges along the way. The most pressing open problems on the analysis side turned out to be:

Challenge 1	Integrate contextual knowledge in the analysis
Challenge 2	Enable communication between metadata extractors

To illustrate Challenge 1, we observe that to find relevant videos about Elon Musk, it helps to know that his rocket project is called SpaceX, and that their rockets are launched from Cape Canaveral in Florida. For Challenge 2, we note that if speech recognition has identified a sound snippet as either Musk, the innovator, or Mūsk, the Iranian city, and face recognition at close time coordinates has identified a face as belonging to either Elon Musk or Drew Houston, then the combination of these strengthen the proposition that the footage is showing Musk.

The solution requires the ability to represent, learn, and apply contextual knowledge in the analysis process. The goal of this project is to use the machine learning paradigm called structured prediction to deal with these challenges. Intuitively, structured prediction is a generalized approach that utilizes prior knowledge of the structure of both the input and the output, whereas in more traditional methods, the focus is almost exclusively on the input. The purpose of this project can be summarized as extending the MICO platform results on holistic search to a system incorporating holistic analysis.

2 Background

A solution to the above-mentioned challenges requires methods to represent, learn and apply contextual knowledge in a multimodal analysis setting. One way to represent previous knowledge, is to use a knowledge base encoding facts as relationships between different entities. The fact that Elon Musk is the founder of SpaceX would for instance be represented by the tuple *(Elon Musk, Founder, SpaceX)*. The aggregation of a knowledge base of sufficient size is in itself a significant task, since the information is typically extracted from loosely structured web sources. Moreover, any automatically generated knowledge base is bound to lack some potentially important information. Therefore, using machine learning to model the entities and relations in the knowledge base, the existence of unobserved links can be predicted. The paradigm of structured prediction is ideal for this task.

Structured prediction is a generalized form of supervised learning, concerned with finding structure in both input and output. Cortes et al. present a new approach for general structured prediction, *Voted Conditional Random Fields* (VCRF), with nice results on part-of-speech tagging [Cor+15], suggesting guarantees of good performance on less data. As VCRF is a general algorithm for structured prediction, the challenges presented can be modeled using VCRF.

Google's Knowledge Vault uses structured prediction (more specifically, link prediction) to learn semantic information implicit to the original data [Don+14]. They use structured prediction on facts from a database of knowledge (also known as a knowledge base). If a fact is not explicitly stored, structured prediction is used to compute the likelihood of the missing fact. A similar use case is IBMs question answering system Watson. Watson was used in 2011 to beat human experts in the quiz show *Jeopardy!* [Fer+10]. In the underlying machinery, Watson used a knowledge base as a part of scoring competing alternative answers with a confidence. Today, Watson is used to assist nurses in cancer treatment [Dod16].

The Umeå research group, with Drewes leading the effort, is currently working on Abstract Meaning Representation (AMR [Ban+13], see Fig. 1 for a minimal example). AMR is a graph representation of sentence semantics that has started to draw significant attention in the research community. A major goal of this research is to devise algorithms for processing AMR graphs as well as learning methods that allow to derive them automatically from input sentences or speech. See, e.g., [Dre17; BD17; Chi+17] for some of the hitherto achieved results.

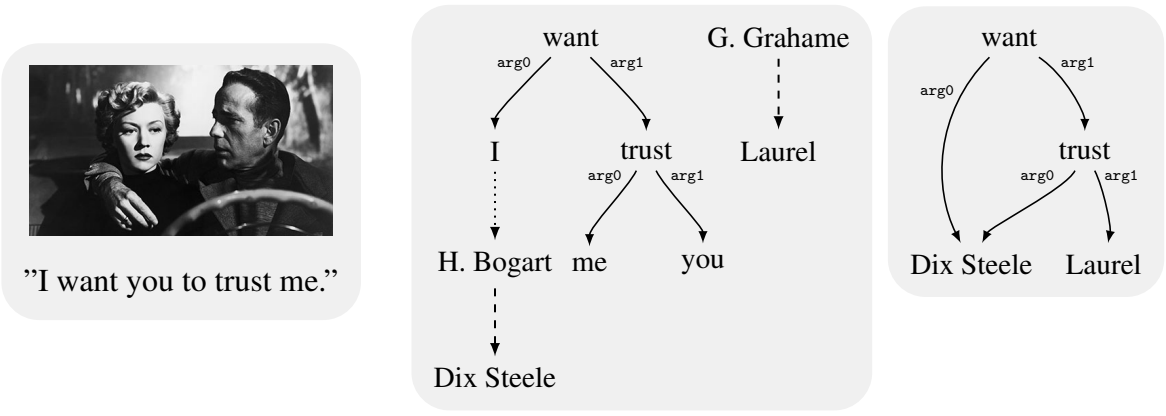

Figure 1: Speech recognition, face detection, and speaker identification are applied to a video (left) and combined with IMDB metadata to derive a set of AMR fragments (middle). The fragments are unified into an AMR describing the scene (right).

3 Project description

The project has three main themes; automatic metadata extraction, semantic modelling, and machine learning. More specifically, the project aims to apply these in a video analysis setting. Metadata extraction allows the computer to see beyond the pixels of the video and make observations of which objects, faces and words appear. Semantic modelling can be used to put all metadata extracted into a structured summary that is better suited for decision making.

3.1 Representing semantic information

A necessary prerequisite for addressing Challenges 1 and 2 is knowledge representation. The developed model must be useful both for encoding and exchanging facts, and must also be applicable at several layers of abstraction, to represent semantic and contextual information. Thus, the first problem to be addressed is to develop such a model.

Since it is infeasible to encode every single piece of information, we also need ways to extract new information from existing knowledge. The second problem to address is thus how new knowledge can be derived from existing data in a provably efficient way. Modelling and learning knowledge are problems which can be solved using structured prediction.

Our initial approach will use knowledge bases as a foundation to represent semantic information. There are large open-source data sets and the general inference problem in knowledge bases has efficient solutions, as described in Section 2. However, the general knowledge base approach needs to be extended in order for media analysis to utilize the information. A framework on top of that used in, e.g., the MICO project or general knowledge bases such as Freebase needs to be developed, allowing video analysis tools to use semantic information.

Research question 1 Develop a framework for metadata extractors to utilize and exchange semantic information, based on open standards such as RDF.

Research question 2 Extend the results of Cortes et al. on VCRF to the inference problem in knowledge graphs.

3.2 Using contextual information in media analysis

The final problem is the incorporation of the knowledge model into the analysis process. There are several media analysis tasks with existing algorithms that involve hypothesis selection (e.g. n-best [CS89]). However, most approaches try to capture this in the internal representation, e.g., implicit semantics captured by neural networks performing speech-to-text analysis. This makes it is necessary to (1) retrain a model if new facts are added or (2) perform online learning. The third option, which we adopt, is to explore algorithms which externalise this knowledge.

Research question 3 Context aware algorithms based on structured prediction will be devised and integrated with the framework described in Section 3.1.

Research question 4 Produce open sourced data sets for training video analysis tools using contextual information.

Research question 5 Linking metadata between media analysis tools during the ongoing analysis.

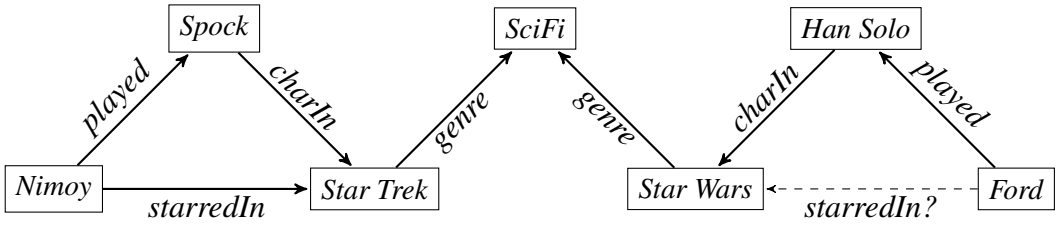

Figure 2: Directed labeled graph constructed from a knowledge base. Predicting the likelihood of the missing link $(\textit{Ford}, \textit{starredIn}, \textit{Star Wars})$ can utilize actors' structural similarities.

3.3 Theory and method

Given input and output spaces $\mathscr{X}$ and $\mathscr{Y}$, a feature function f to be maximised, and an input $x \in \mathscr{X}$, structured prediction is the task of finding the best output y^*: The input x could be a sentence or an image and the output a part-of-speech tagging or image segmentation. The input and output space can be arbitrarily complex. Such a multidimensional data set calls for sophisticated algorithmic approaches. As a result, many approaches for structured prediction algorithms rely heavily on learning theory results. One such example is the work by Cortes et al. on *voted conditional random fields*, mentioned in Section 2.

The structured prediction problem for knowledge bases can be illustrated using Figure 2. A set of facts (represented as tuples such as $(\textit{Nimoy}, \textit{played}, \textit{Spock})$) can be modelled as a graph where edges represent relationships between entities. The left triangle containing *Nimoy*, *Spock*, *Star Trek* can be used to recover the missing edge between *Ford* and *Star Wars* based on structural similarities. This problem is commonly referred to as *link prediction*. Traditionally, large amounts of data to train on is necessary. For our purposes the open source Freebase data set is an interesting starting point, as it contains 637 million non-redundant facts on 40 million entities and 35,000 relation types [Don+14]. Freebase is also stored in RDF, allowing integration with previous work on the MICO project.

3.4 Time plan

The student will follow the General Study Plan for doctoral studies in Computing Science, which includes courses and seminar series. The most important milestone is the student earning his or her PhD degree. Approx. 2 years into the project the student will publish a licentiate thesis as an intermediate step. Activities towards these goals are: (i) writing a report of an initial literature study, (ii) visiting a partner research group for knowledge exchange at the beginning of the project, and (iii) publish scientific articles addressing the research questions described above, at a rate of 2–3 conference and journal articles per year. Work on the research questions 2 and 3 will commence during the first year (2018), work on questions 1 and 4 commence in 2019 and research question 5 will be fully addressed in 2020.

4 Preliminary and expected results

The expected results of this project are two-fold. Firstly, it will provide a practical framework for enriching media analysis with semantic information, thus extending the MICO platform. Secondly, it contributes to establishing a new generation of media analysis algorithms, with the main focus on video analysis. Such algorithms require a solid theoretical foundation, as a continuation of the work presented by Cortes et al. Preliminary results include our above-mentioned work on AMR, MICO, and adaptions of the link prediction problem for knowledge bases to the VCRF algorithm. In the spring of 2017, Johanna Björklund supervised a master thesis project [Dah17] on using the VCRF algorithm for link prediction on knowledge bases.

5 Team and collaborators

The host company Codemill AB has 10 years of experience in delivering tailored video-asset management solutions to, among others, ITV, SAT1, and the Guardian News and Media. The industrial advisor Dr. Johanna Björklund was named in 2016 as one of Sweden's 10 most innovative entrepreneurs, and by DI Digital in 2017 as one of Sweden's most promising entrepreneurs under 40. She is the founder of three companies in digital video, and an active researcher in the field of Algorithmics.

The active involvement of the academic advisor Prof. Frank Drewes guarantees that the project reaches the scientific depth required to successfully complete the candidate's PhD studies. Drewes's broad array of scientific collaborators include groups in Canada (University of Western Ontario), Germany (Universities of Bremen, Dresden, Gießen, Leipzig, and Munich), Great Britain (University of Edinburgh), Italy (Universities of Padova and Pisa), The Netherlands (Leiden University), South Africa (Stellenbosch University), and the USA (Universities of Southern California and Notre Dame). The recent collaboration on semantic analysis with Shay Cohen at the University of Edinburgh will be of particular value, giving access to a leading expert in Machine Learning and opening the door one of Europe's leading institutions in the field of semantic analysis. A valuable asset for the project is the research group's close collaboration with the Zooniverse group at the University of Oxford, responsible for the platform Zooniverse.org, a tool for crowdsourcing human-generated classification of images [Mas+16].

References

[Aic+15] P. Aichroth, E. Berndl, J. Björklund, et al. "MICO – Media in Context". *Proc. 2015 IEEE International Conference on Multimedia & Expo, EU Projects Papers*. 2015.

[Ban+13] L. Banarescu, C. Bonial, S. Cai, et al. "Abstract Meaning Representation for Sembanking". *Proc. 7th Linguistic Annotation Workshop, ACL 2013 Workshop*. 2013.

[BD17] J. Blum and F. Drewes. "Language Theoretic Properties of Regular DAG Languages". *Information and Computation*. 2017.

[Chi+17] D. Chiang, F. Drewes, D. Gildea, et al. "Weighted DAG Automata for Semantic Graphs". 2017.

[Cor+15] C. Cortes, P. Goyal, V. Kuznetsov, et al. "Kernel Extraction via Voted Risk Minimization". *Proc. 1st Workshop on Feature Extraction: Modern Questions and Challenges*. 2015.

[CS89] Y.-L. Chow and R. Schwartz. "The *N*-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses". *Proc. Speech and Natural Language*. 1989.

[Dah17] A. Dahlgren Lindström. "Structured Prediction using Voted Conditional Random Fields". Master thesis. Umeå University, Department of Computing Science, 2017.

[Dod16] G. Doda. *Manipal Hospitals Announces National Launch of IBM Watson for Oncology*. *IBM news*. 2016.

[Don+14] X. Dong, E. Gabrilovich, G. Heitz, et al. "Knowledge vault: a web-scale approach to probabilistic knowledge fusion". *The 20th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, New York, NY, 2014*. 2014.

[Dre17] F. Drewes. "DAG Automata for Meaning Representation". *Proc. 15th Meeting on the Mathematics of Language*. 2017.

[Fer+10] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, et al. "Building Watson: An Overview of the DeepQA Project". *AI Magazine*. 2010.

[Mas+16] K. Masters, E. Oh, J. Cox, et al. "Science learning via participation in online citizen science". *Journal of Science Communication*. Apr. 2016.

[TG16] S. Tsekeridou and P. Gros. *Technical Review Report: Media in Context, Grant no. 610480. Period 3/3*. 2016.