Bachelor thesis, Institute of Computer Science, Freie Universität Berlin

Human-Centered Computing (HCC), AG NBI

Towards interpretability in unsupervised NLP

– Exposé –

Tim Korjakow

tim.korjakow@campus.tu-berlin.de

Supervisor: Jesse Josua Benjamin

Berlin, May 8, 2019

1 Motivation of the thesis

Research on Explainable Artificial Intelligence, often called XAI, is currently wildly distributed and characterized by several competing ideas and approaches from a vast number of fields including computer science, mathematics, social sciences and philosophy. Research is mostly focused on explainability in computer vision since the inner workings of a model can directly be translated into a graphic representation. In contrast to that development stands the fact that most of humanity's knowledge is encoded in text and XAI algorithms and methods from computer vision most often cannot be directly applied to results of NLP algorithms. Therefore there is an urgent need to research these methods in the context of NLP.

In Project IKON, we are developing a data-driven application at a major natural history research institution in order to make potentials for knowledge transfer between research projects and society actionable. To this end, it features a data visualization of semantic relations between research projects supplemented by links to infrastructures (e.g., collections or labs) and knowledge transfer activities (e.g., workshops or lectures). To generate the semantic relations, we have developed a Topic Modelling Pipeline with a Singular Value Decomposition at its heart. It is thought that this method shares a limited amount of expressiveness with different other linear methods due to its purely linear nature [AHM+17]. Additionally, when considering how we can make the results of our pipeline more interpretable, we, as creators of the system, encountered significant doubts over the meaningfulness of approaches such as parameter manipulation or choosing between algorithms for dimensionality reduction [BMG18] in our use case.

In short, the fundamental challenge of interpretability in Project IKON is: which model can we use that is potentially more interpretable, and how precisely can what aspect be made interpretable for humans in order to support the context of identifying potentials for knowledge transfer at the research institution? The latter illustrates a significant gap in the related work, as contextuality (both considering the way in which the output of a machine learning algorithm is operationalized in an algorithmic system as well as the situated context of use) is a sorely neglected aspect of interpretability [Mil17].

2 Related work

With the surge of the application of machine learning (ML) systems in our daily life there is an increasing demand to make operation and results of these systems interpretable for people with different backgrounds (ML experts, non-technical experts etc.). Contrary to these efforts, interpretability as term has become an ill-defined objective [Lip16] for research and development in ML algorithms since there is no widely agreed upon definition of it.

Miller et al. [MHS17] supports this point by conducting a literature study and uncovering that interpretability research is rarely influenced by insights from the humanities, especially connected fields as explainability or causality research.

This thesis builds upon these concepts and tries to transfer their critical insights into the sub field of unsupervised natural language processing - an often overlooked discipline in the context of interpretability research.

3 Goal setting

As formulated in the introduction, the main problem for project IKON is uncertainty about the interpretability of the existing model. This thesis should therefore examine existing interpretability techniques, their applicability to the existing model or a contending one and subsequently a decision for one of them. The chosen model will be implemented and augmented by a number of interpretability techniques. The results will be assessed with the help of an exemplary, qualitative user test with an expert user from the natural history museum. This goal setting directly leads to the following set of questions:

- Which techniques to enhance interpretability of models are out there and how are they characterized?
- How can the existing topic modeling pipeline be augmented or changed in order to enhance different aspects of interpretability?
- Does the usage of these methods result in an enhanced understanding on the end users side?

4 Procedure and methods

One of the most crucial parts of the existing topic modelling pipeline is the document embedding. Currently a simple information retrieval method - Tf-Idf - is used, but the main drawback of the technique is that it, as all Bag-of-Words method, completely disregards word order and semantic dependencies in texts. Since that may be a integral part especially in scientific literature a more complex model is needed which is able to capture this dependencies. In order to do this a short summary of the state of research of document embeddings is presented and an contender to the existing NLP pipeline will be chosen.

In order to gain a reproducible overview over the status of XAI research in the field of NLP a literature mapping study according to Petersen et al. [PFMM08] is going to be conducted. This should result in a number of papers which are, according to the process, good representatives of the literature base and therefore also of current research efforts.

A quantitative and qualitative analysis of these papers should summarize occurring XAI methods and categorize them according to proposed criteria e.g. Liptons's "Properties of Interpretable Models" [Lip16] or Robnik-Sikonja's criteria [RB18].

In this analysis a model is more interpretable if it supports more interpretability techniques that were sourced from the literature analysis. Therefore the next step involves checking which techniques are applicable to the two contending models. Based on the

absolute counts the more interpretable model is chosen. Following a number of supported techniques are selected and going to be applied to the chosen model.

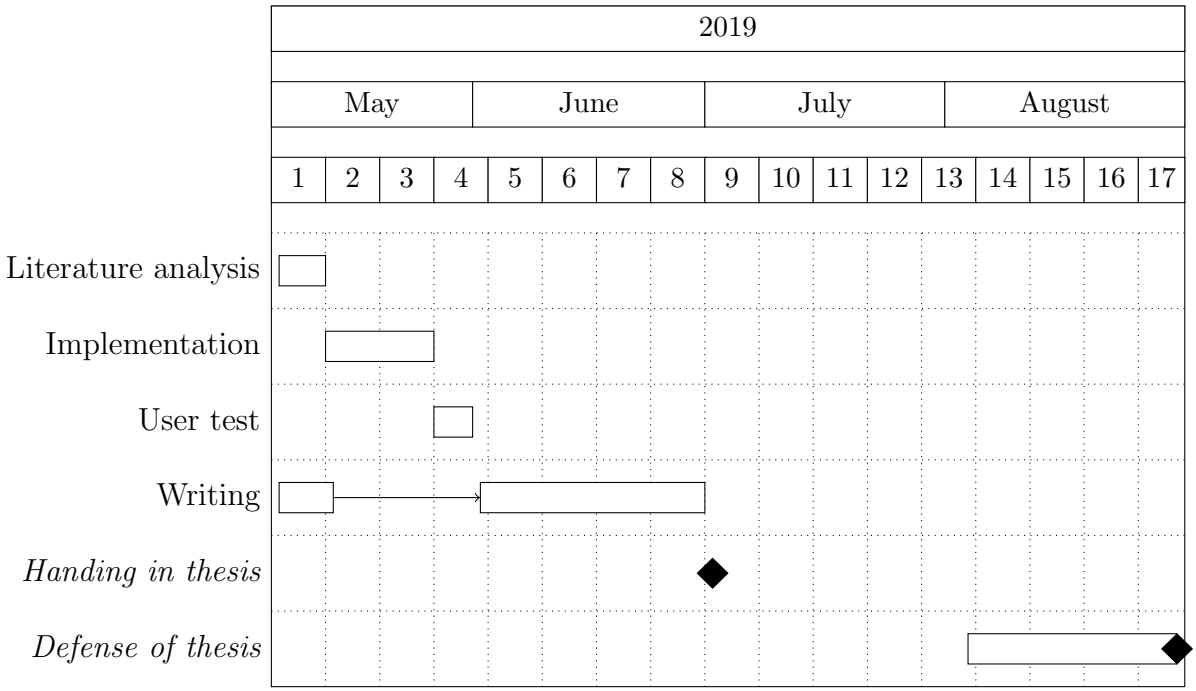
As an exemplary analysis of the impact of these techniques on the human understanding a qualitative interview with domain experts from the natural history museum will be conducted. The qualitative test, which is inspired by the often used method of Think Aloud Tests, should generate insightful information and pointers for further research.

5 Implementation

One of the main technical challenges and parts of the implementational work will be the augmentation of the current topic modelling pipeline by a document embedding technique. Since the performance of the model greatly depends on this step, it is crucial to have well learned vector representations of the document base. Currently there is a corpus of circa 114000 scientific documents available in order to train the model. If that is not enough to gain expressive document embeddings, one may include pretrained word embeddings via e.g. BERT [DCLT18] to introduce external information into the model and enhance the semantic coherence of the learned embeddings. This path should be taken with caution since it is connected to an hardly determinable amount of complexity. Research in the field of transfer learning for document embeddings is still in its infancy.

In order to adhere to the current research and industry standards, the implementation of this thesis is going to be done in Python. Based on the chosen model which is going to be augmented with explainability mechanisms further packages and technologies are going to be selected.

6 Time calculation



References

[AHM+17] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLOS ONE*, 12(8):e0181142, August 2017.

[BMG18] Jesse Benjamin, Claudia Müller-Birn, and Rony Ginosar. Transparency and the Mediation of Meaning in Algorithmic Systems. October 2018.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October 2018.

[Lip16] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, June 2016.

[MHS17] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*, December 2017.

[Mil17] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*, June 2017.

[PFMM08] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic Mapping Studies in Software Engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swindon, UK, 2008. BCS Learning & Development Ltd.

[RB18] Marko Robnik-Sikonja and Marko Bohanec. Perturbation-Based Explanations of Prediction Models. pages 159–175. June 2018.