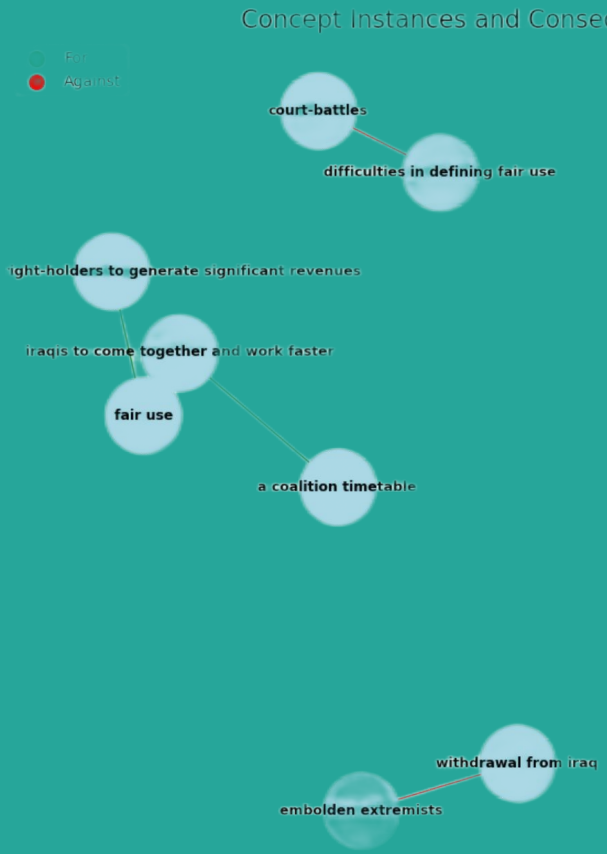
XAI/BERT Update

Argumentation Knowledge Graph Construction

for Ue AI and Visualization on December 12th, 2024

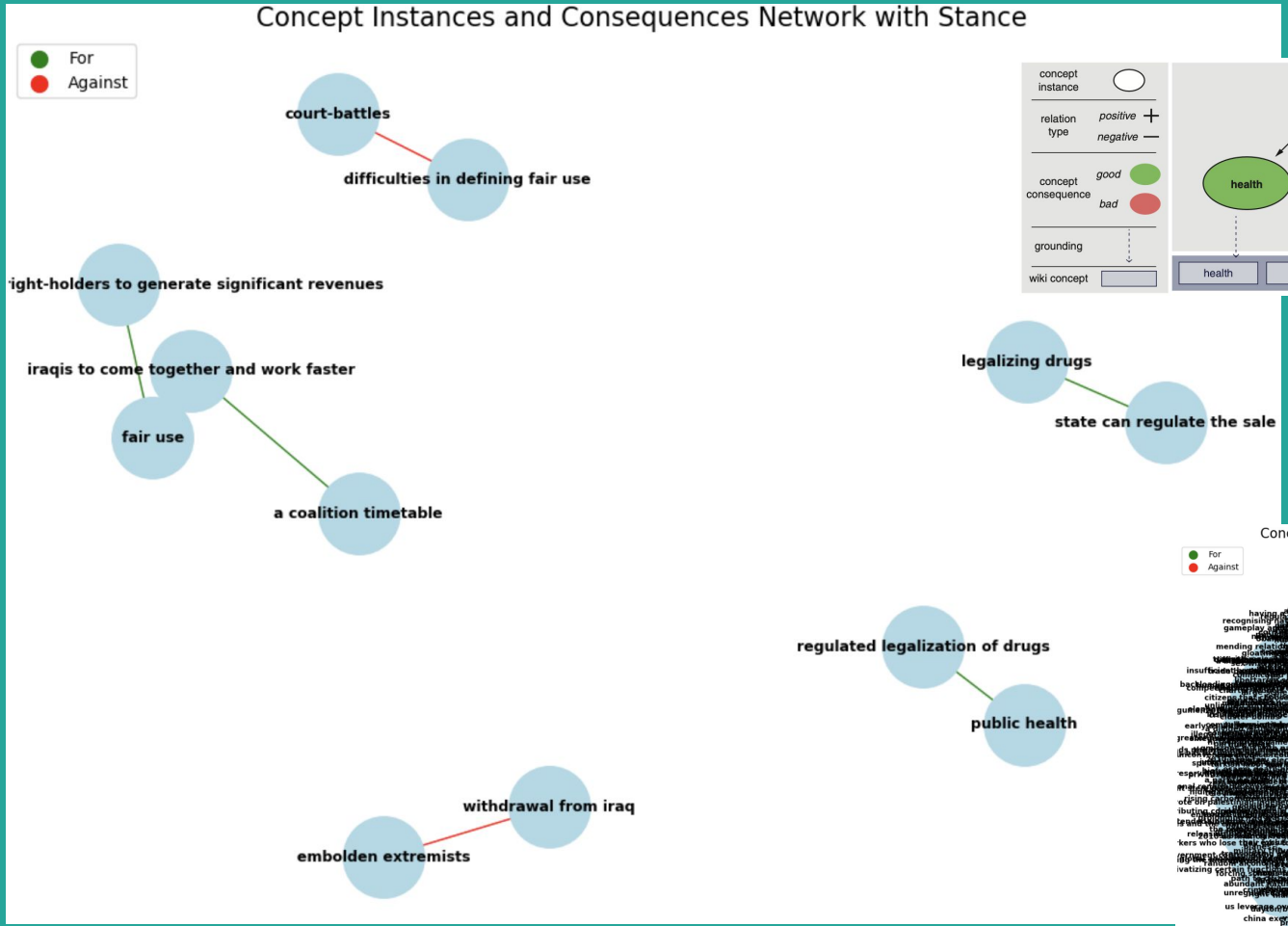
@JKU Linz

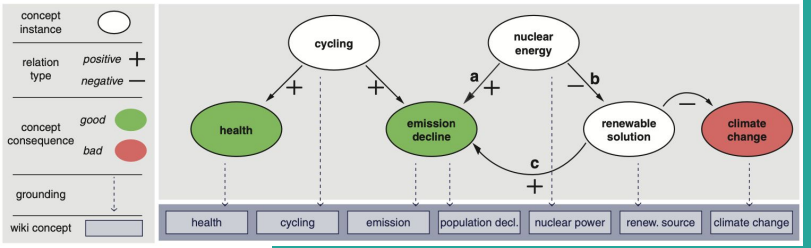
Qais Almanasra
Chhitiz Buchasia
Felix Eichhorn
Jack Heseltine



Knowledge graphs to explain a wide range of argument mining tasks

1st Step: **Graph visualization** (xAI #0)

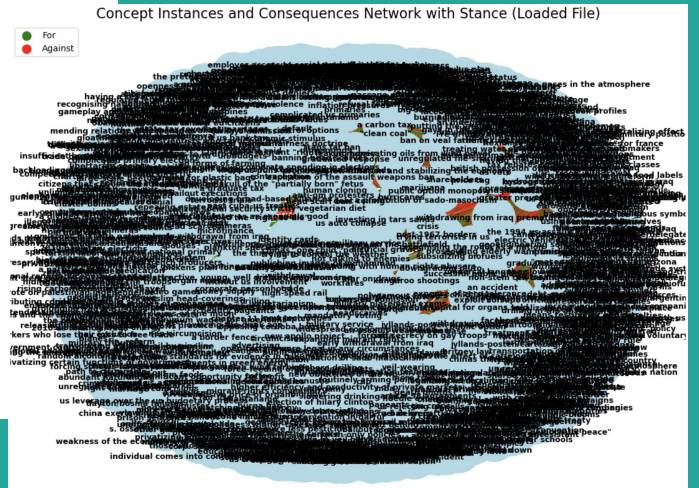




Above: from the paper

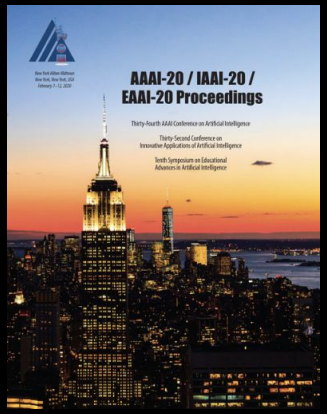
Left: local graph

Bottom: global graph (entire dataset)



Continued: **Graph visualization** (xAI #0)

(We'll be interested in the relation types, i.e. the edges between concepts)



Paper of Interest (AAAI-20)

The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

End-to-End Argumentation Knowledge Graph Construction

Khalid Al-Khatib,[1] Yufang Hou,[2] Henning Wachsmuth,[3]
Charles Jochim,[2] Francesca Bonin,[2] Benno Stein[1]
[1]Bauhaus-Universitat Weimar, Germany
[2]IBM Research, Ireland
[3]Paderborn University, Germany

Our Approach

(our story)

Curiosity-led Project

Researcher contact ->

+ found a way to make a current contribution, too, by recreating the setup using BERT

- "Guess you probably can achieve better performance with small encoder transformer models (e.g., BERT)"
- "The original idea is to leverage the scheme defined in the paper and the corresponding *knowledge graphs to explain a wide range of argument mining tasks*"
- "Unfortunately we haven't had the chance to touch on the explainability part in the follow up work"

Dataset (Features + Example)

1.	Input.Stance	For,
2.	Input.Claim	"By legalizing drugs, the state can regulate the sale",
3.	Input.Tagme_concepts	"{'regulation', 'state (polity)', 'recreational drug use', 'legalization
4.	Input.Babelfy_concepts	{'drug'},
5.	Answer.rel	Relation,
6.	Answer.concept_1	legalizing drugs,
7.	Answer.concept1.entities	Legalization|drug,
8.	Answer.effect	Positive,
9.	Answer.concept_2	**state can regulate the sale,**
10.	Answer.concept2.entities	state (polity)|regulation,
11.	Answer.GoodBad	Good,
12.	Answer.concept_3	"citizens, public health, state, government, public safety, taxes, budge
		sales"

Focus

- **Input**: a sentence
- **Output**: No/Relation (0/1)
- But think about it: **this is hard.**
  We are talking about language
  - *Lexical features*
  - *Syntax features*
  - *Sentiment features*
  - *Semantic features*

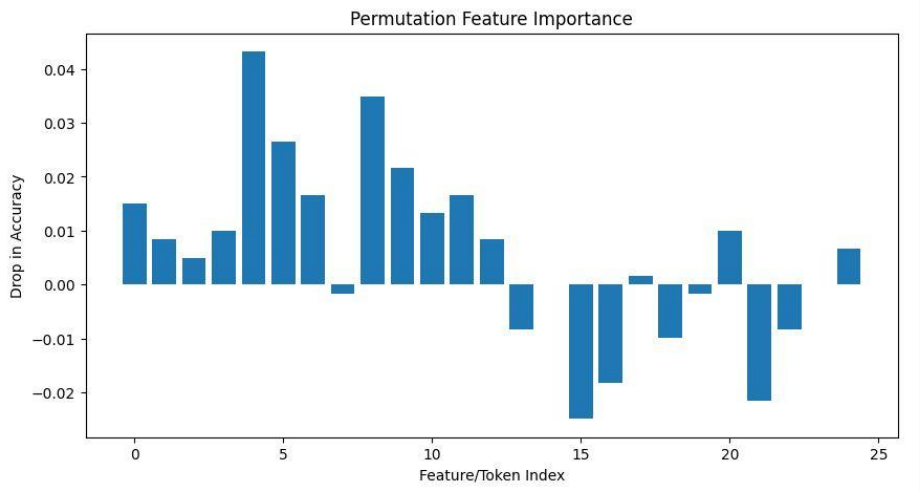
- Focused on the **Relation Detection** part, using BERT
- Using the annotated dataset contributed by the original paper
- Applied XAI approaches to just the Relation Detection
- Hoping to understand robustness and overall quality of BERT in this use case

Goal: We want to understand robustness of language input as well as BERT as intermediary repr.

- **... language:**
  - *Lexical features*
  - *Syntax features*
  - *Sentiment features*
  - *Semantic features*

#1: Permutation Feature Importance (1)

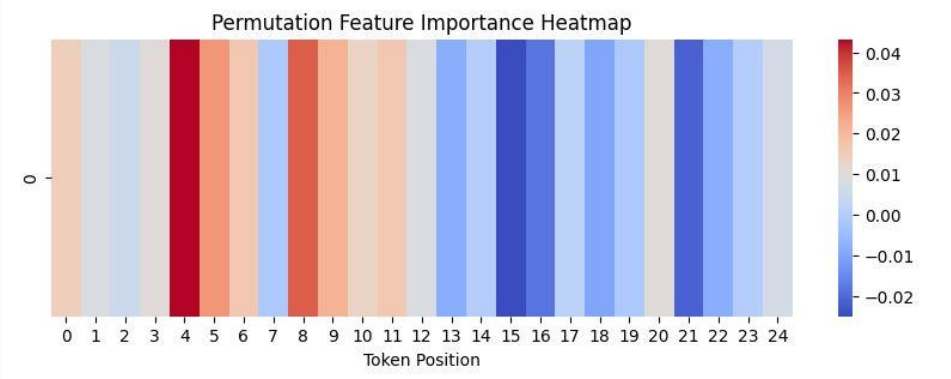
- Identify impactful input features by permuting them and measuring accuracy drop
- model is robust, accuracy drops between ~0.05 and ~-0.05
    - first half of tokens has a noticeable effect on accuracy => important features
    - second half less impact, likely due to redundancy and padding

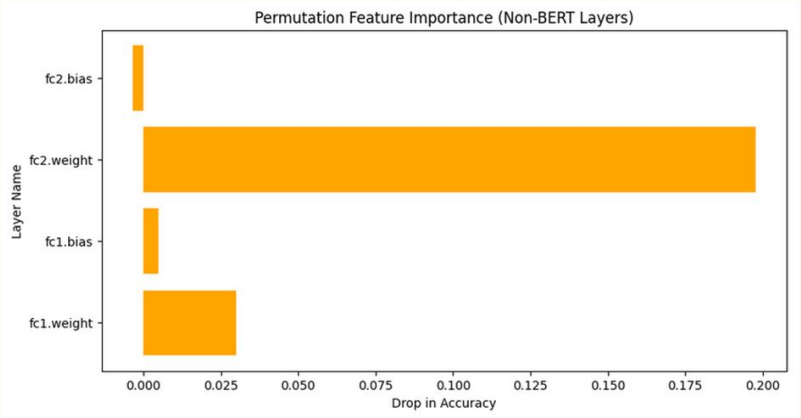


#1: Permutation Feature Importance (2)

- Heatmap highlights the contribution of initial tokens and the decreasing impact of later ones

- Layer Analysis:
  - Second fully connected layer has the highest impact, maps representations to output probabilities





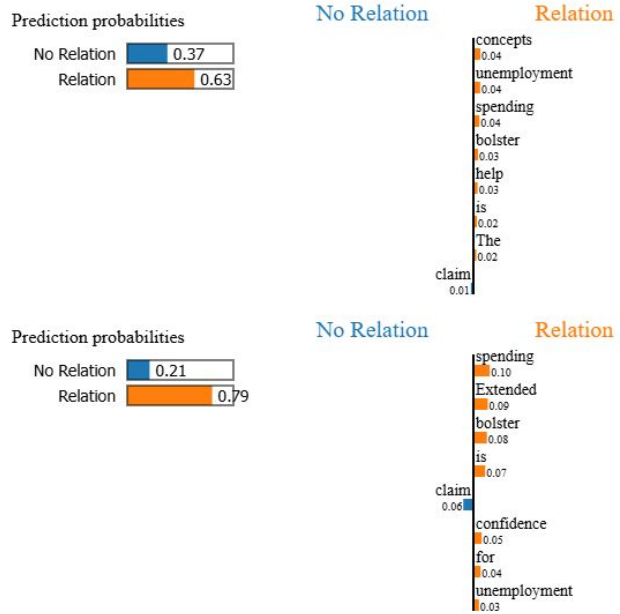

#2: Local Interpretable Model-agnostic Explanations (LIME) for text (1)

- Look at individual predictions of two models and try to understand how they work differently.

- Later features also important for individual predictions contrary to observations in previous methods.

#2: LIME for text (2)

```
text: The claim "Extended unemployment benefits help bolster confidence/spending" is for the concepts "unemployment benefits or unemployment benefits"
true label 1
prediction 100: 1
prediction 500 1
```



Text with highlighted words

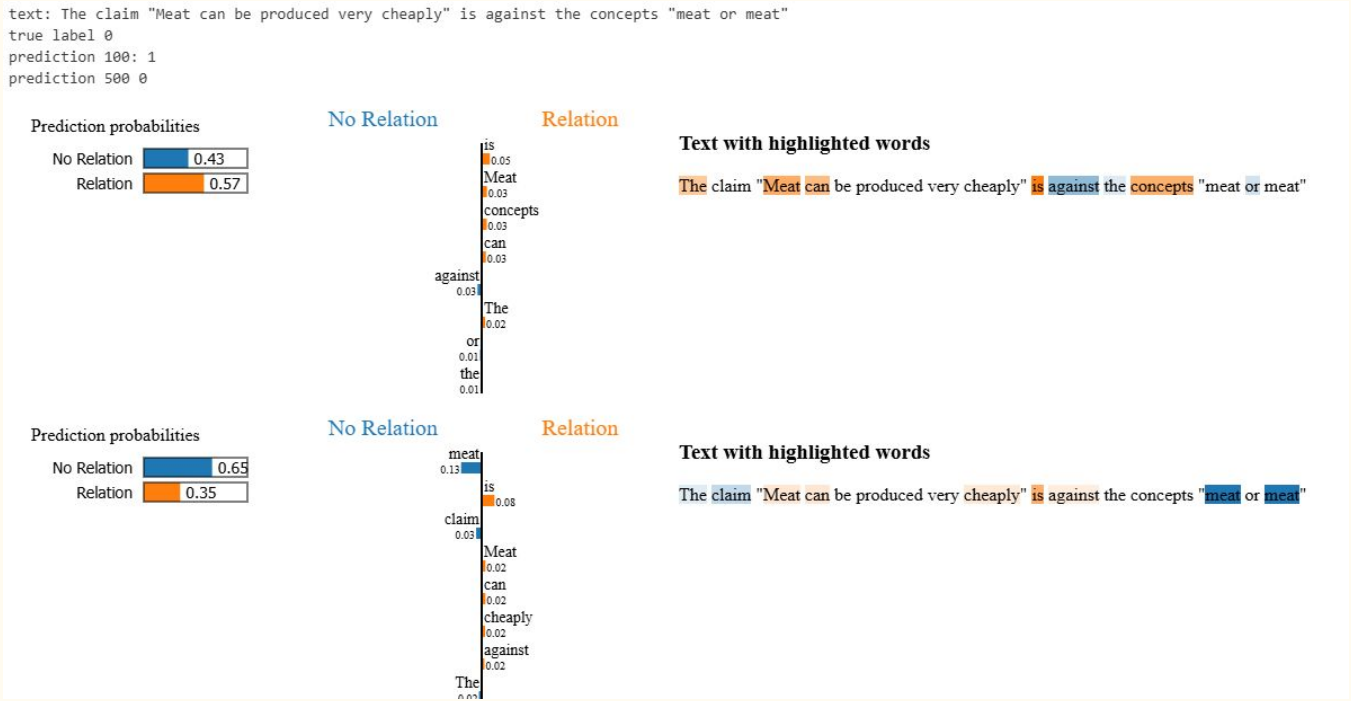
The claim "Extended unemployment benefits help bolster confidence/spending" is for the concepts "unemployment benefits or unemployment benefits"

Text with highlighted words

The claim "Extended unemployment benefits help bolster confidence/spending" is for the concepts "unemployment benefits or unemployment benefits"

#2: LIME for text (3)

- Every prediction that we looked at had both the models giving importance to words that we added to create the sentences.



Idea: Words (Lexicon) -> Concepts ~ Semantics

- *Semantic features*

#3: Concept Bottleneck Models (CBMs, 1)

Input.Claim > BERT > c.1

c.1 > own classifier > c.2 sub-step

c.2 > own classifier > task y

(here: **stance prediction**)

We have fully annotated dataset including the concepts, so can take advantage of this

Potential Future Work: Testing with Concept Activation Vectors (TCAV) for **c.1**

We were motivated to understand the importance of concepts in our use of BERT in our pipeline

- Using the intermediary concept classification **c.2 sub-step**
- One-hot classification proved to be challenging (concept imbalance in dataset possibly)
  - `Wind energy comes naturally from the environment.,"republicanism, olympic games, high tech, seattle, internment"`
  - `Vegetarian foods have as many health risks as animal foods.,"republicanism, olympic games, high tech, seattle, internment"`


#3: Concept Bottleneck Models (CBMs, 2)

Input.Claim > BERT > c.1

c.1 > own classifier > c.2 sub-step

c.2 > own classifier > task y

(here: **stance prediction**)

=> Once we figured out concept predictions (and in any case feeding the true labels we had as well) we could get better performance on stance-prediction using (true) concepts (especially)

Performance improvement with concepts

- Bias toward *against* persists but is reduced with concepts
- Predicted concepts are noisier but effective as well (surprisingly)
- actual concepts perform better overall.

(Story) Conclusion (1)

5-Ws/How: This approach enabled us (model creators, who) to understand better the relationship between concepts and their consequences (what) for the textual dataset before model design (when) for feature selection and feature creation (where) which was otherwise not easy to understand/use (why) using a network based graph (how), or even the non-visualized input.

- We moved from words to concepts, later in our process
- Finding that while model shows robustness w.r.t. Input –
- — Concepts are especially rich for mining additional representations to solve a prediction task as in our task-prediction case

(Story) Conclusion (2)

Potential Data Set Improvements

- The **dataset does not comprises of many examples for the same concept**, it would be nice to have more data for all the individual concepts in the dataset. This will enable to do more concept related analysis/training.

Future Work
(**end** of our story?)

We want to talk more to the 2nd researcher who advised us and who now happens to be teaching in Linz about our results + see if there are any research questions we want to investigate

Thank you.

Slides/presentation: Jack Heseltine
Project repo:
https://github.com/jku-icg-classroom/assignment-2-model-explanations-cube5