Connecting the Dots: NLP support for Analyzing Arguments, Evidence and Translations in the Humanities

Acknowledgements to Dr. Johannes Daxenberger and Chris Stahlhut, M.Sc. for the slides

TECHNISCHE UNIVERSITÄT DARMSTADT

Hamburg,
21. October 2019

Humanities Research Needs Search

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Knowledge-Based Search: Retrieving Translations

2. Information Retrieval and Classification: Arguments

3. Interactive Machine Learning: Evidence Detection

How to Computationally Approach Extinct Languages: A Case Study on Hittite

TECHNISCHE UNIVERSITÄT DARMSTADT

- BMBF-funded "**Ce**ntre for the **Di**gital **Fo**undation of **R**esearch in the Humanities, Social, and Educational Sciences"
- Coordination: TU Darmstadt (Iryna Gurevych), partners: Goethe-Universität Frankfurt a.M. and Leibniz Institute DIPF Frankfurt a.M.
- CEDIFOR intends to contribute to bridging the gap between research in the humanities and computer based methods, and help researchers to master the characteristic problems in this process

SPONSORED BY THE

Federal Ministry of Education and Research

Parts of this work are accepted to be published at the DHd conference 2017:

Daxenberger, J., Görke, S., Siahdohoni, D., Gurevych, I., & Prechel, D. (2017). Semantische Suche in Ausgestorbenen Sprachen: Eine Fallstudie für das Hethitische. In Proceedings of the DHd 2017 (p. to appear).

CEDIFOR Pilot Project: Building a Framework for Semantic Search on Cuneiform Scripts

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Goal:** Make the translations of Hittite cuneiform documents more accessible by adding an additional layer of information
- **Methods:** Lexical-semantic methods based on Natural Language Processing technology for a deeper exploration of the documents

- *Cooperation with Prof. Doris Prechel and Susanne Görke, Altorientalische Philologie, Johannes Gutenberg-Universität Mainz*

JOHANNES GUTENBERG UNIVERSITÄT MAINZ JG|U

The Cuneiform Script

TECHNISCHE UNIVERSITÄT DARMSTADT

- Cuneiform Script (4 cent. BC) is documented on clay tablets only
- Well-Documented Cuneiform Languages include
  - Akkadian (Semitic)
  - Sumerian (Isolated)
  - Hittite (Indo-European)

Digitizing Cuneiform Clay Tablets...

7	--	[... GÙ]B-laza [...]
7	A	Vs. II 6' [... GÙ]B-la-za (Rasur) Vs. II 7' [...]
8	--	[... Ḫ]I.A adanna [...]
8	A	Vs. II 7' [... Ḫ]I.A a-da-an-na Vs. II 8' [...]
9	--	[nu=tta DUGKA.G]AG gulšantan [...]
9	A	Vs. II 8' [... DUGKA.G]AG gul-ša-an-ta-an Vs. II 9' [...]
10	--	[... walh]it? [...]
10	A	Vs. II 9' [... wa-al-h]i?-it Vs. II 10' [...]

Digitizing Cuneiform Clay Tablets...

7	--	[... li]nks [...]
8	--	[...] zu Essen [...]
9	--	[... dir] den eingeritzten [Bech]er [...]
10	--	[...] mit [walḫ]i-Getränk$^{?}$ [...]
11	--	Ich ließ [...]
12	--	[Iss Erstklassiges]
13	--	[und] stille deinen Hunger!
14	--	[Trink Erstklassiges]
15	--	[und] stille deinen Durst!
16	--	[...]
17	--	[Der Bedrü]kte aber [...]
18		[...] komm [wieder] in Ordnung!

Motivation: Accessing the Content of Digitized Cuneiform Documents

TECHNISCHE UNIVERSITÄT DARMSTADT

- Digitalization and Accessibility are highly desirable goals
- Currently no means to automatically process the transliterations
- But we can work with the translations (which are a bit special...)

- **Semantic Search**
  - Assume we have only a rough idea of what we want to search/find, e.g. a king, a god or a city
  - There are plenty of named entities in the document falling into these categories

- **Primary requirements for the search engine**
  - Return both exact matches as well as results motivated by semantic relatedness
  - Intuitive usage by non-experts

Current Data

TECHNISCHE UNIVERSITÄT DARMSTADT

- Original documents: Hittite and Akkadian
  - about 700 documents and growing…
  - approx. 400.000 translated words
- Translations in German, Italian, French, and English
- Content: rituals, myths, state treaties, prayers
- Parallel Text: Transliterations and Translations
  - Mapping on the level of meaning units
- Title and introductory comments for each document

Particularities of the Data

TECHNISCHE UNIVERSITÄT DARMSTADT

- Comments: in-line annotations and footnotes
- Mixed dialects/languages
- Normalized translations vs. text evidences
- Language experts/translators have diverging conventions (e.g. meaning of special characters)
- Damaged and thus fragmentary text (to the extent of illegibleness of entire documents)

Indexing: Technologies

TECHNISCHE UNIVERSITÄT DARMSTADT

- DKPro Core/UIMA as convenient NLP processing framework
  - Reader recognizes document structure and parallel translation/transliterations (e.g. footnotes)
  - Tokenizer: OpenNLP and Language Tool
  - Fixing token boundaries due to fragmentary text
  - POS-tagging: OpenNLP
  - Lemmatizing: MateTool, LanguageTool
  - Word Sense Disambiguation: DKPro WSD/Lesk
  - Semantic Preprocessing: Uby
  - Writer: MySQL writer (12 tables)

https://dkpro.github.io/

Indexing: Semantic Preprocessing

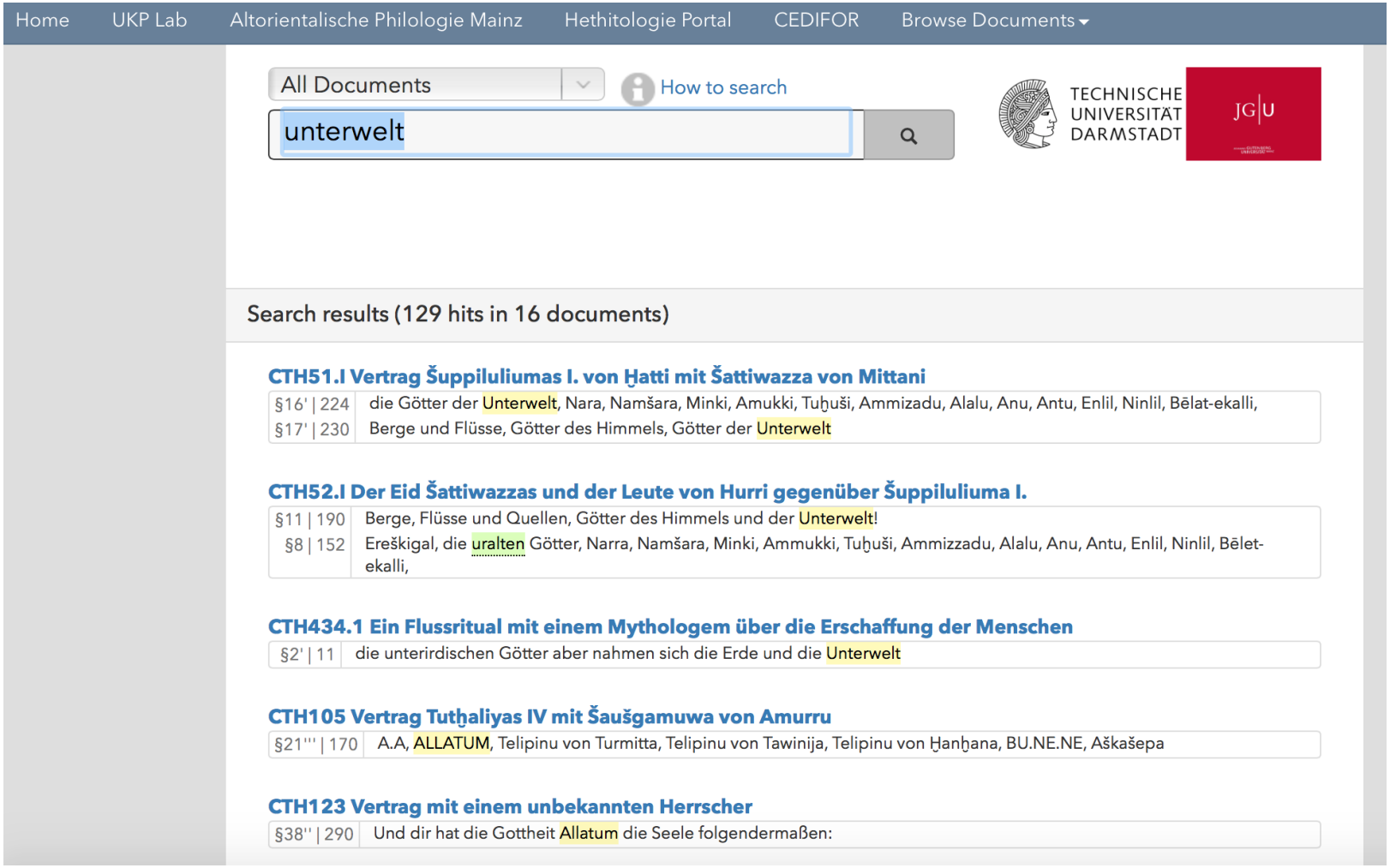
TECHNISCHE UNIVERSITÄT DARMSTADT

- Uby (combines sense information from WordNet, FrameNet, VerbNet, Wiktionary, Wikipedia and others)
  - **Semantic Labels**
    - Semantic Field (e.g. person)
    - Category (e.g. fruit)
    - Domain (e.g. sports)
  - **Synonyms**

- **Custom Dictionaries** (extensible)
  - Alternative Spellings
    - e.g. *Ḫattuša, Hattusa, Hattuscha, Hattuşaş, Hattusha, Hatusha, ...*
  - Hypernyms
    - e.g. *Herrscher, Großkönig, König : Pitḫana, Anitta, Labarna, Ḫattušili, ...*

https://dkpro.github.io/dkpro-uby/

Query Interface



Computationally Approaching Extinct Languages: Lessons Learned

TECHNISCHE UNIVERSITÄT DARMSTADT

- Typical issues when computationally approaching extinct languages:
  - **Small data** (not just small training data)
  - **Non-standard vocabulary** (translations)
  - **Fragmentary** text

- Some solutions
  - Implement **language-independent**, broad approaches
  - any available **meta-information** (commentaries, footnotes, etc.) should be used
  - Use **NLP machinery off-the-shelf** a far as possible, add custom solutions where required

Humanities Research Needs Search

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Knowledge-Based Search: Retrieving Translations

2. Information Retrieval and Classification: Arguments

3. Interactive Machine Learning: Evidence Detection

Argument Retrieval and Classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Topic

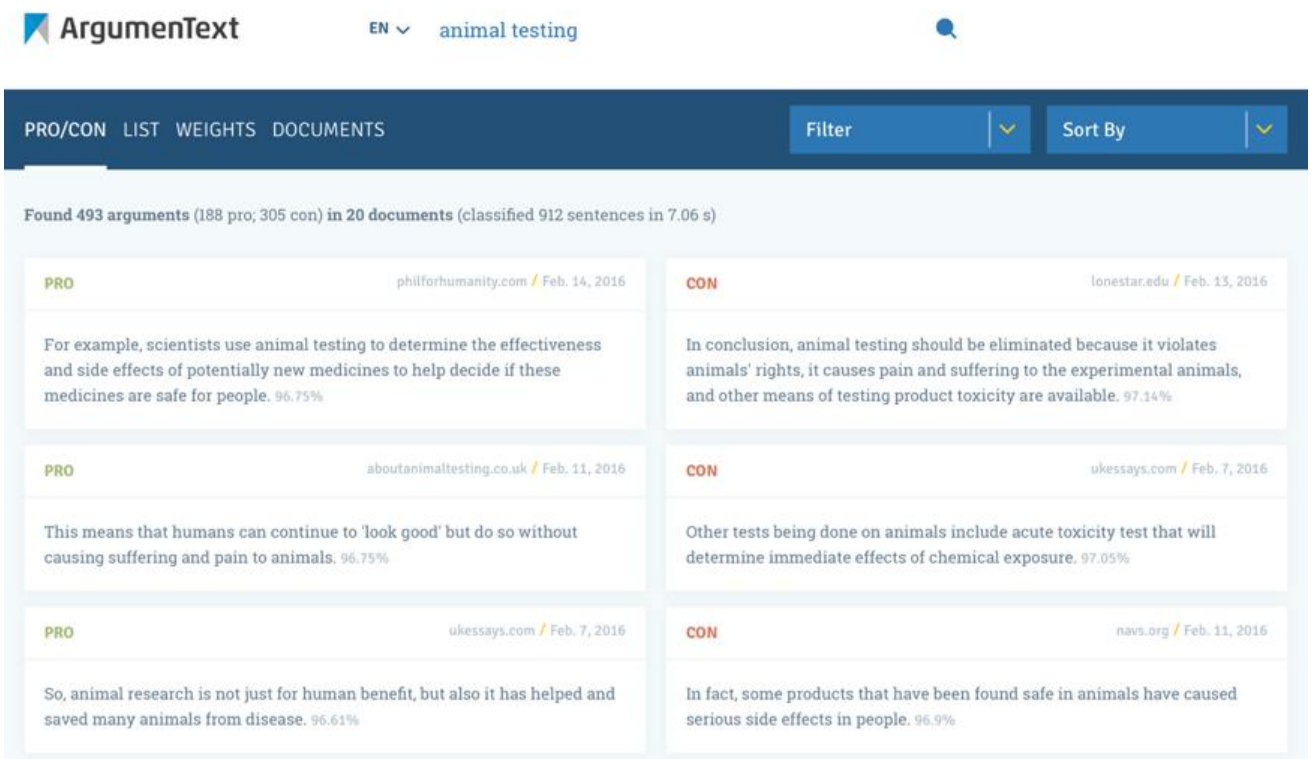
Autonomous trucks

PRO: The use of such technology will reduce fatal truck and bus accidents by 20 per cent, resulting in 800 less deaths on European roads.
http://www.theautochannel.com

CON: These components are still expensive today; for example, the 600-rpm spinning light-imaging radar system that crowns most current autonomous vehicles costs upward of $70,000.
http://www.strategy-business.com

Search Engine: Retrieval and Classification

TECHNISCHE UNIVERSITÄT DARMSTADT



UKP

Challenges in the Web Scenario

TECHNISCHE UNIVERSITÄT DARMSTADT

Heterogeneity	Scalability	Robustness
Multiple text types	Small training resources	Specific domains
Different languages	Expert annotations	Various applications

Training Data Generation

TECHNISCHE UNIVERSITÄT DARMSTADT

Data

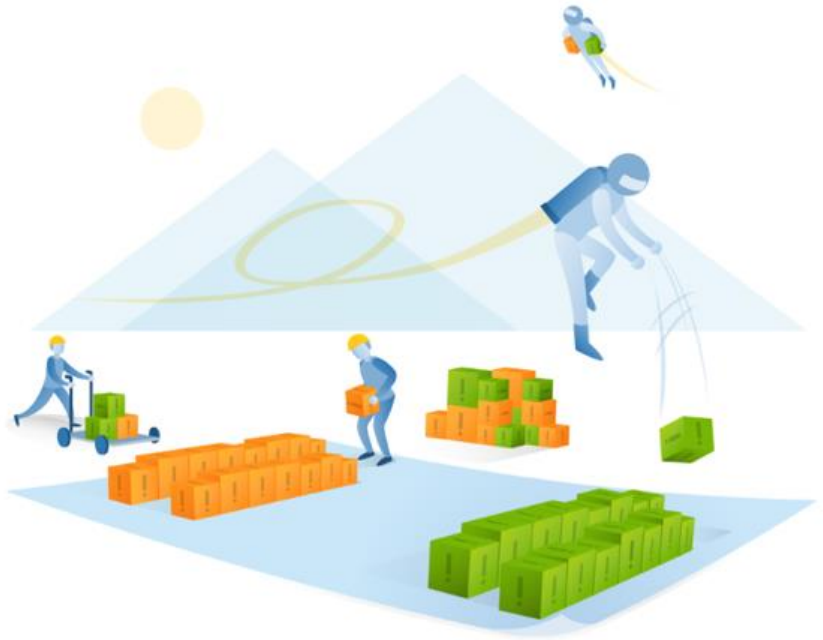
- Heterogeneous text types
- Collected from web searches
- 100+ topics

Task

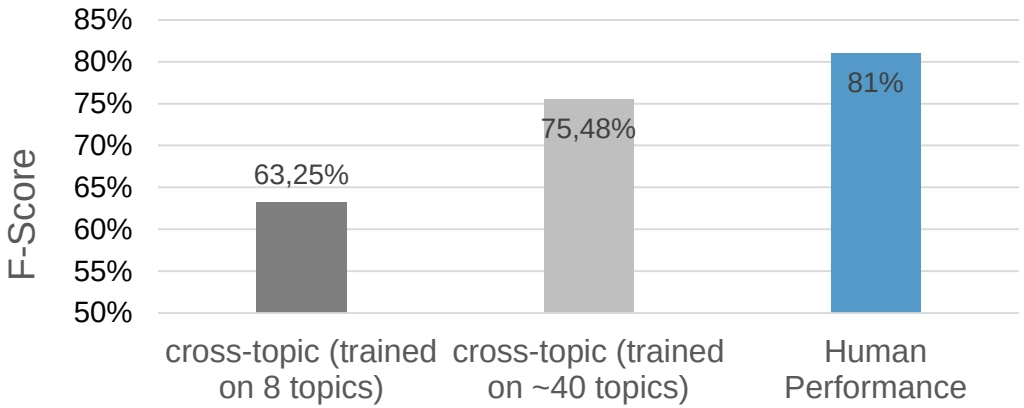
- Given a (controversial) topic
- Label text as pro or con argument

Crowdsourcing

- Amazon Mechanical Turk



Results (BERT): Stance + Argument Detection



Reimers, N., Schiller, B., Beck, T., Daxenberger, J., et al. (2019)

- 3-class problem: pro, con, non-argument
- Approx. 5pp improvement over BiLSTM

Visualizing Argument Clusters

Topic: Abortion

- cluster167: *Fetuses are incapable of feeling pain when most abortions are performed.*
- cluster63: *Abortion is the killing of a human being, which defies the word of God.*
- cluster91: *Allowing abortion conflicts with the unalienable right to life recognized by the Founding Fathers of the United States.*

Humanities Research Needs Search

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Knowledge-Based Search: Retrieving Translations

2. Information Retrieval and Classification: Arguments

3. Interactive Machine Learning: Evidence Detection

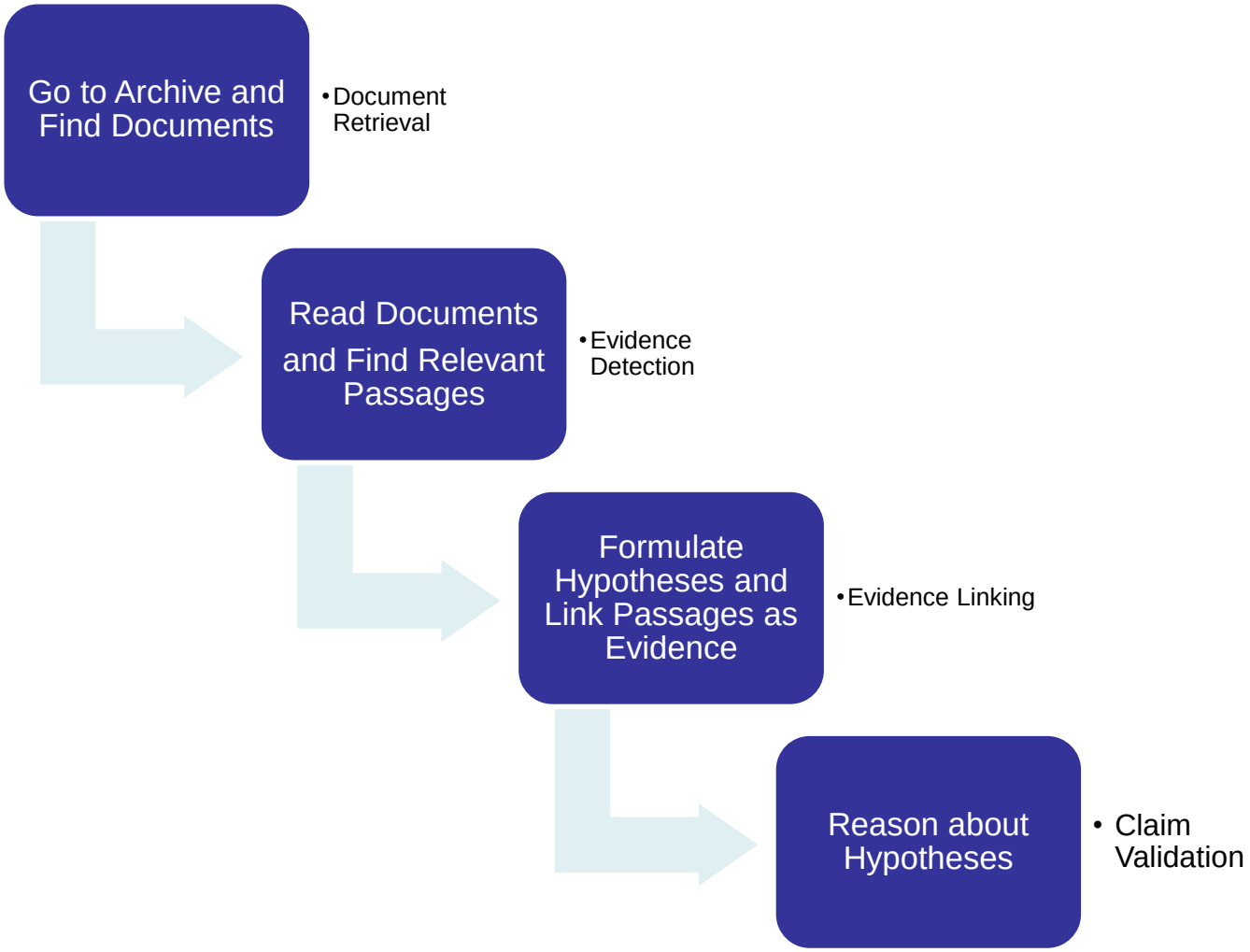
Interactive Evidence Detection for Historical Research

TECHNISCHE UNIVERSITÄT DARMSTADT

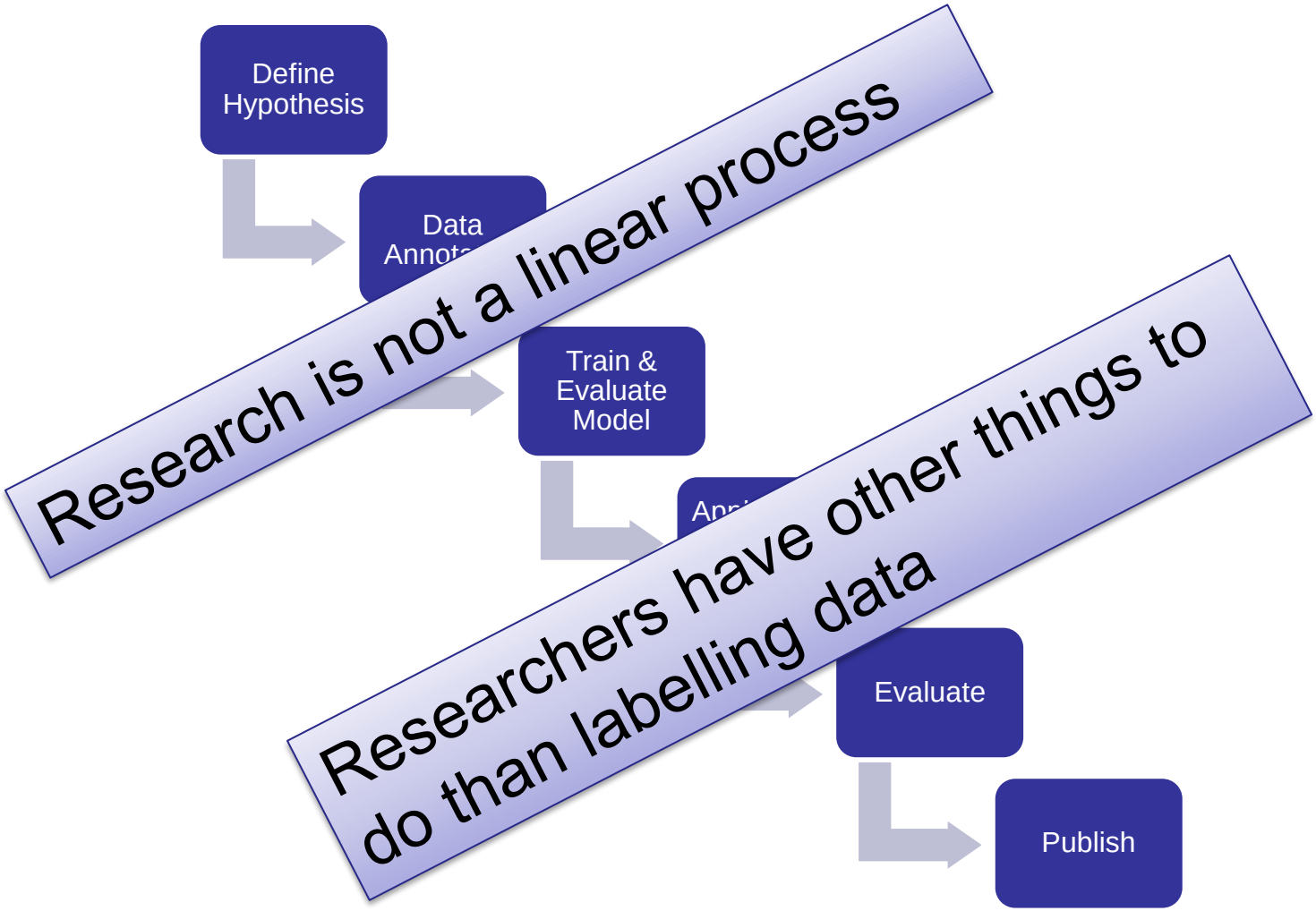
- Historians spend a vast amount of time reading to find the piece of evidence they can use to (in)validate a hypothesis

- **Our goal:** given a research hypothesis, automatically detect and link evidence

IR and Classification as a Viable Solution?



IR and Classification as a Viable Solution?



How do researchers in the humanities and social sciences actually work?

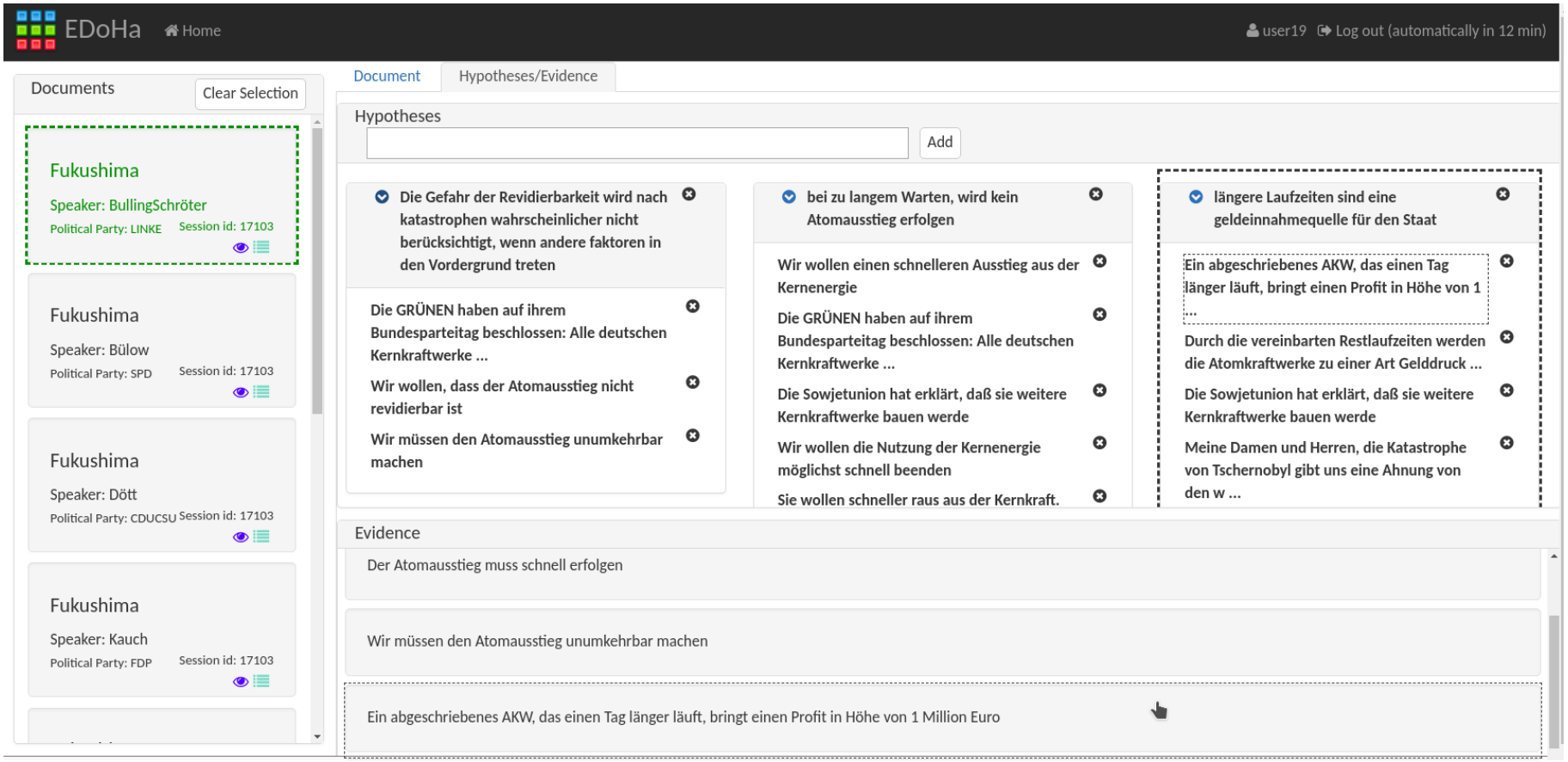
TECHNISCHE UNIVERSITÄT DARMSTADT

- Conducted a user study in a student seminar on the topic of *Environmental catastrophes in the second half of the 20th century*
- 24 students participating
- Students were asked to *Compare the political discourse after the Chernobyl and Fukushima catastrophes*
- To formulate hypotheses and (in)validate them with evidence found in the provided text
- Text sources were 9 speeches from the German parliament
  - 4 from 1986
  - 3 them 2011

- We logged and analysed their behaviour

EDoHa: Evidence Detection fOr Hypothesis vAlidation

TECHNISCHE UNIVERSITÄT DARMSTADT



UKP

Findings

TECHNISCHE UNIVERSITÄT DARMSTADT

- No uniform approach to evidence detection and linking
  - Some users work in different phases of *evidence detection* and *hypothesis validation*, other do not
  - Some users validated multiple hypotheses at the same time, while others validated only one hypothesis at a time.
- No agreement on evidence for similar hypotheses
- Hypotheses with similar evidence show no similarity

  - **-> Solution:** Interactively trained machine learning models

The interactively trained models out-perform BERT in recall

TECHNISCHE UNIVERSITÄT DARMSTADT

	Macro values across both classes			Evidence only		
	F1	Precision	Recall	F1	Precision	Recall
			ED-ACL-2014			
$\text{bilstm}_{\text{direct}}$	0.509 (0.033)	0.514 (0.028)	0.526 (0.039)	0.117 (0.058)	0.091 (0.055)	0.183 (0.053)
$\text{bilstm}_{\text{fine}}$	0.481 (0.043)	0.518 (0.018)	**0.553 (0.047)**	**0.139 (0.064)**	0.088 (0.045)	**0.373 (0.118)**
BERT	**0.540 (0.052)**	**0.590 (0.055)**	0.538 (0.048)	0.118 (0.098)	**0.238 (0.105)**	0.094 (0.096)
			ED-EMNLP-2015			
$\text{bilstm}_{\text{direct}}$	**0.572 (0.062)**	0.566 (0.050)	0.613 (0.075)	**0.225 (0.133)**	0.176 (0.114)	0.340 (0.160)
$\text{bilstm}_{\text{fine}}$	0.544 (0.063)	0.553 (0.046)	**0.631 (0.089)**	0.212 (0.132)	0.145 (0.101)	**0.453 (0.212)**
BERT	0.550 (0.060)	**0.596 (0.084)**	0.558 (0.081)	0.143 (0.118)	**0.251 (0.169)**	0.143 (0.171)

Especially when computational resources are scarce, it is better to use small amounts of in-domain data to train a model:

BERT requires a GPU <-> A bilstm can be trained on a laptop

Humanities Research Needs Search

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Knowledge-Based Search: Retrieving Translations

2. Information Retrieval and Classification: Arguments

3. Interactive Machine Learning: Evidence Detection

Thank you!

TECHNISCHE UNIVERSITÄT DARMSTADT

UKP