Towards Real-World Natural Language Processing

Heike Adel
Bosch Center for Artificial Intelligence

WiDS 2022

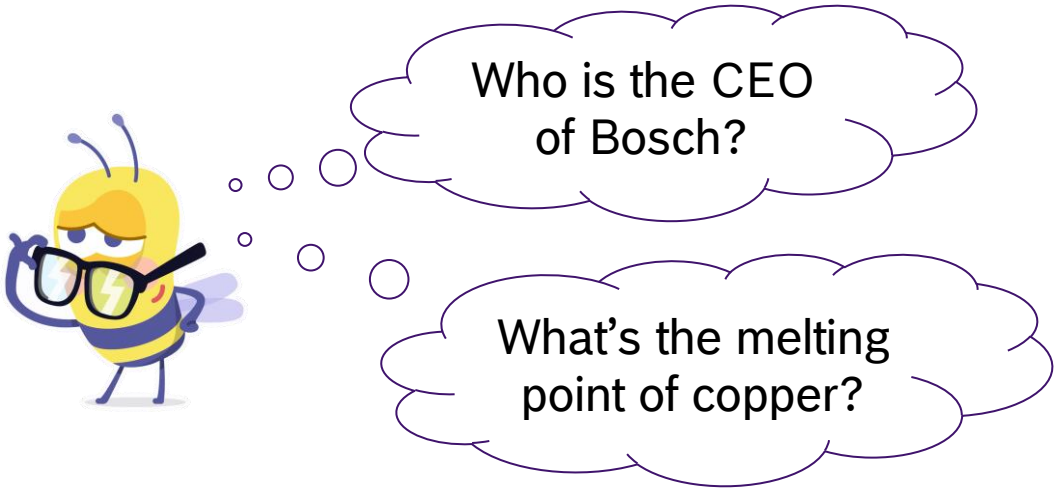
BOSCH

Natural Language Processing
What is this?

- NLP = automatic processing of language

Natural Language Processing
What is this?

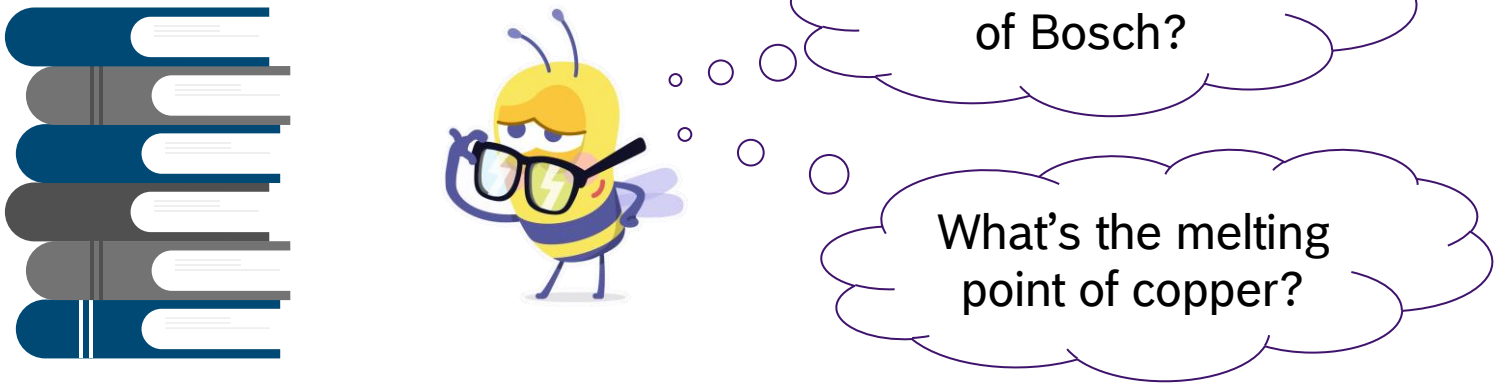
- NLP = automatic processing of language



Heike Adel | 06.07.2022

Natural Language Processing
What is this?

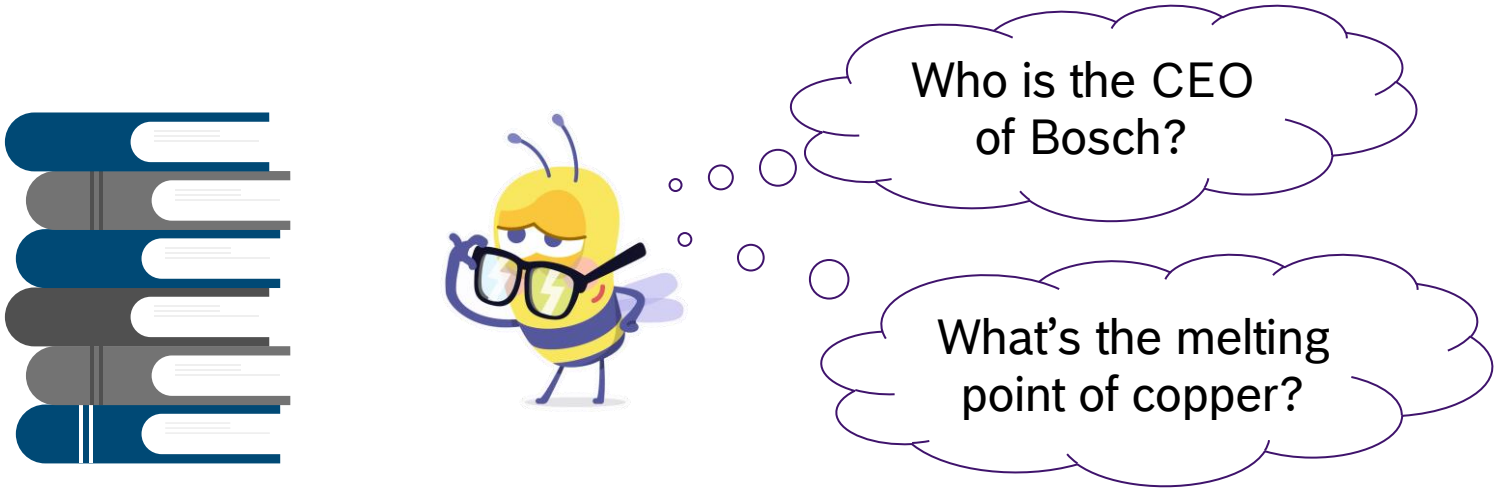
- NLP = automatic processing of language



- Example NLP tasks relevant for answering those questions:
  - Information retrieval (keyword-based search)
  - Information extraction (concept and relation extraction)
  - Question answering

Natural Language Processing
What is this?

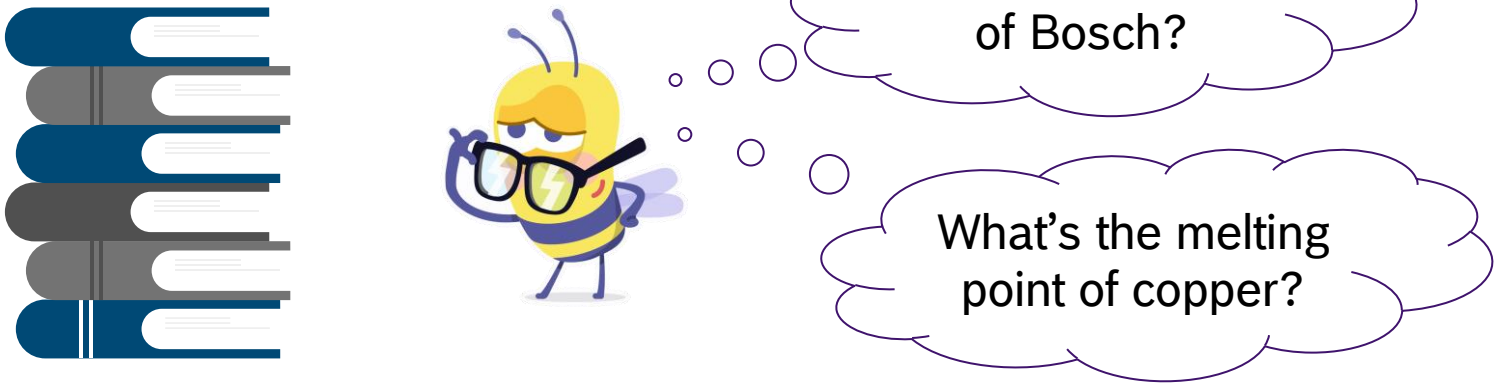
- NLP = automatic processing of language



- Example NLP tasks relevant for answering those questions:
  - Information retrieval (keyword-based search) => limited reasoning abilities
  - Information extraction (concept and relation extraction)
  - Question answering

Natural Language Processing
What is this?

- NLP = automatic processing of language



- Example NLP tasks relevant for answering those questions:
  - Information retrieval (keyword-based search) => limited reasoning abilities
  - Information extraction (concept and relation extraction)
  - Question answering

focus of this talk

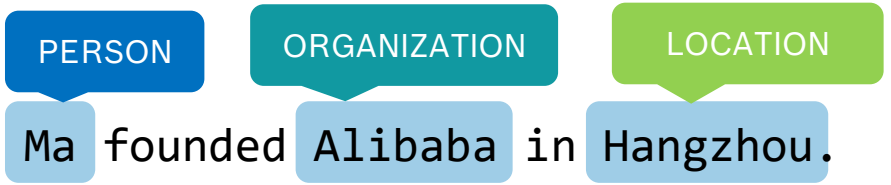
Natural Language Processing
Example Tasks

- **Information extraction** (concept and named entity recognition)
  - Goal: detection and typing of concepts and named entities
  - Example:

```
Ma founded Alibaba in Hangzhou.
```

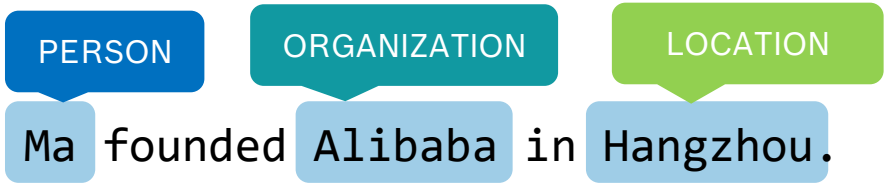
Natural Language Processing
Example Tasks

- **Information extraction** (concept and named entity recognition)
  - Goal: detection and typing of concepts and named entities
  - Example:

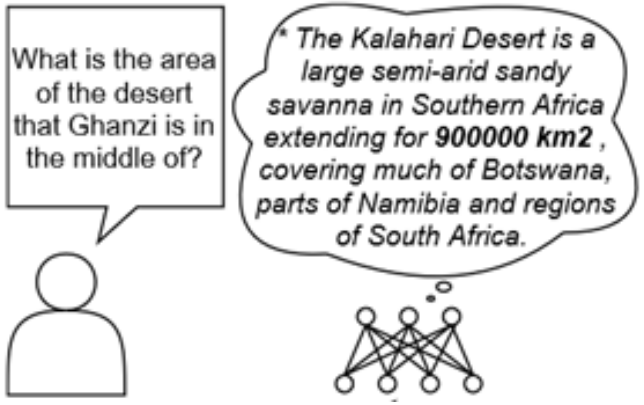


Natural Language Processing
Example Tasks

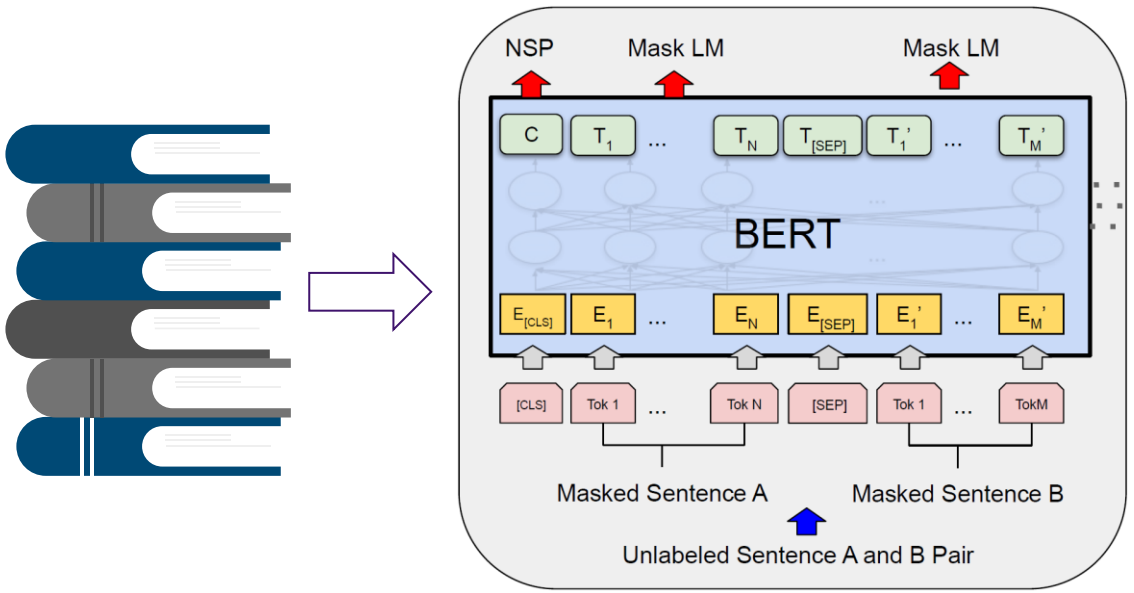
- **Information extraction** (concept and named entity recognition)
  - Goal: detection and typing of concepts and named entities
  - Example:



- **Question answering**
  - Goal: answer questions based on text document(s)
  - Variants: answer-span extraction vs. multiple-choice questions
  - Example:
    Given Wikipedia, find the answer to the question
    “What is the area of the desert that Ghanzi is in the middle of?”

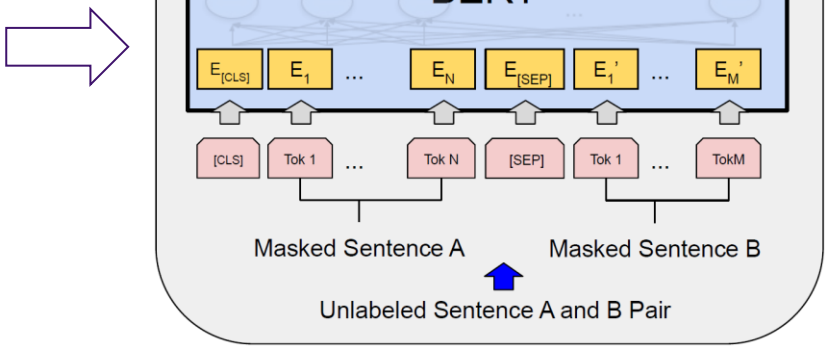


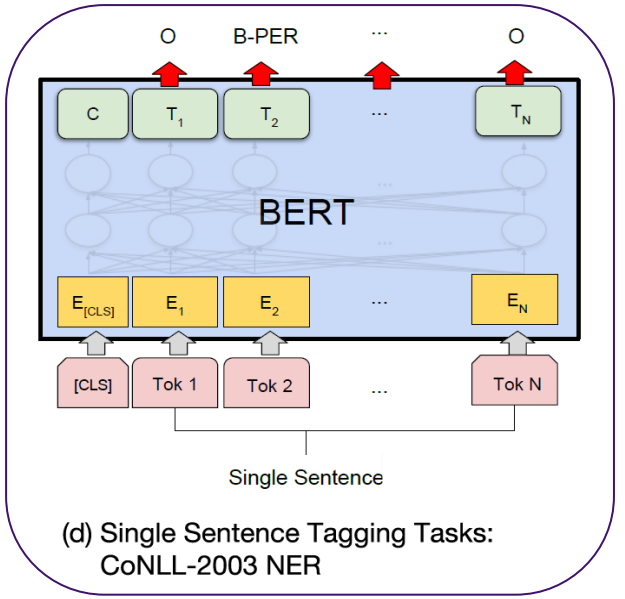
Natural Language Processing
Current State of the Art

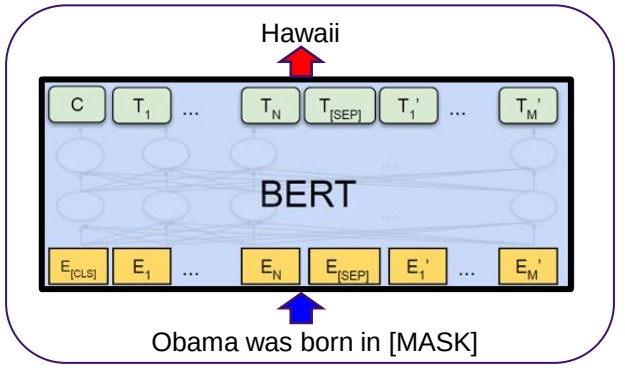


Step 1: pre-train (transformer-based) language model on large amounts of text

Natural Language Processing
Current State of the Art



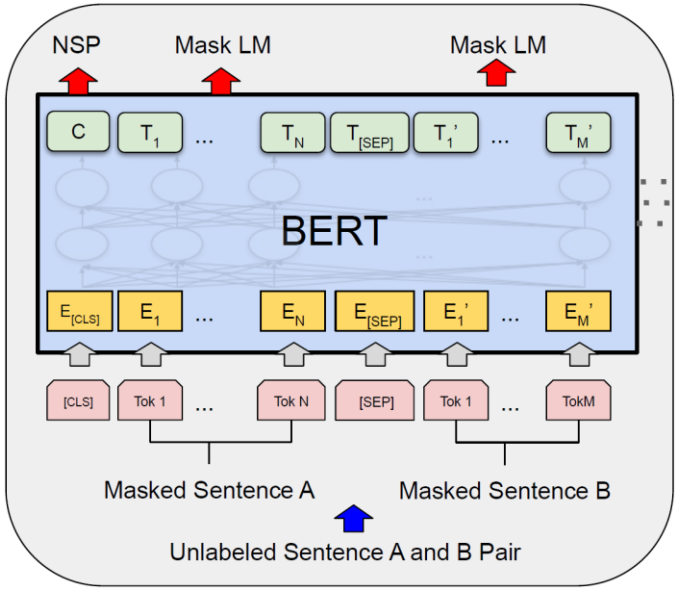


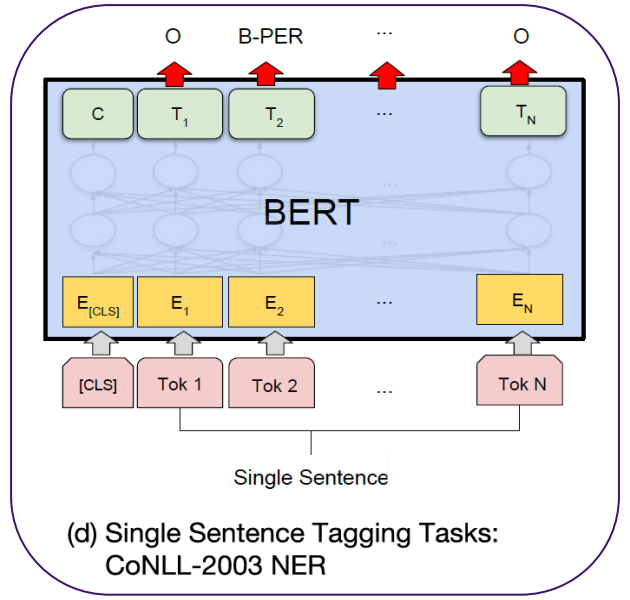


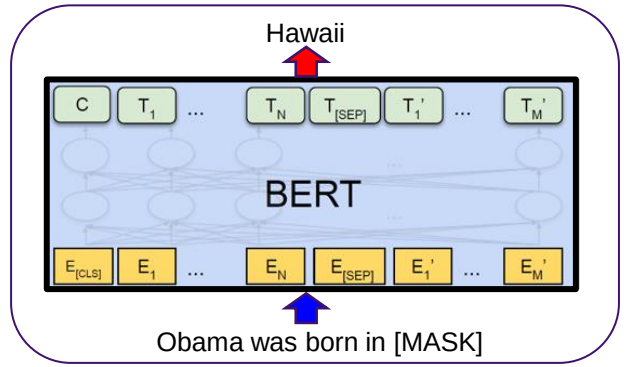
Step 1: pre-train (transformer-based) language model on large amounts of text

Step 2: fine-tune model on task
or: use prompting

Natural Language Processing
Current State of the Art



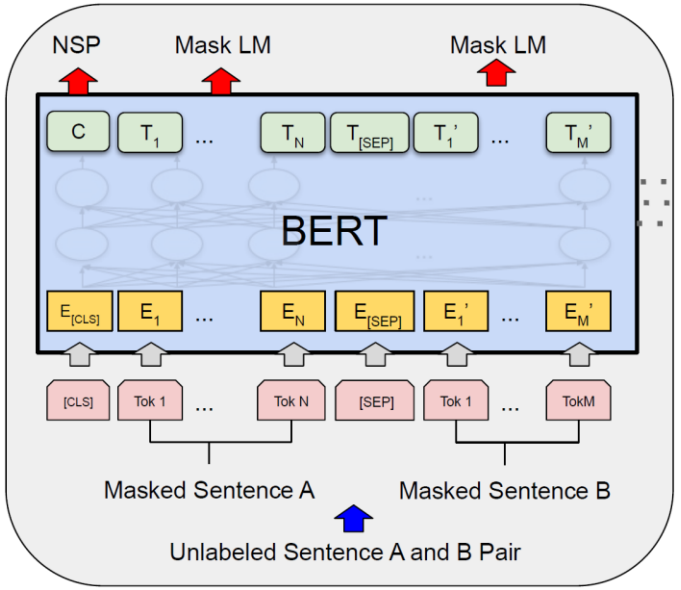


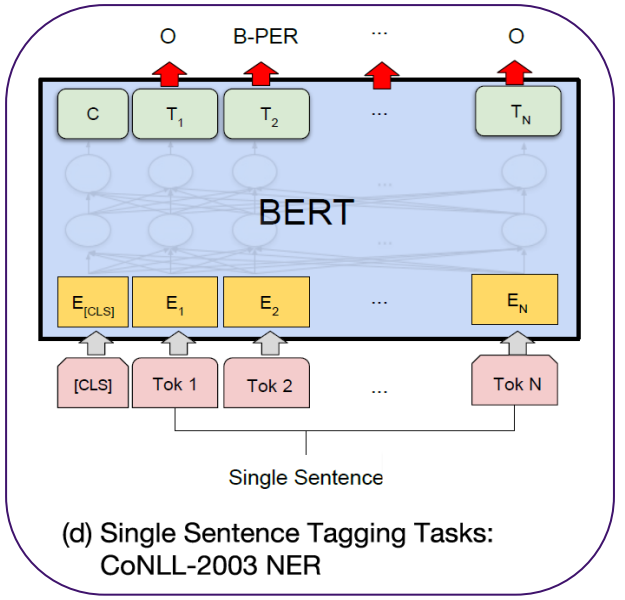


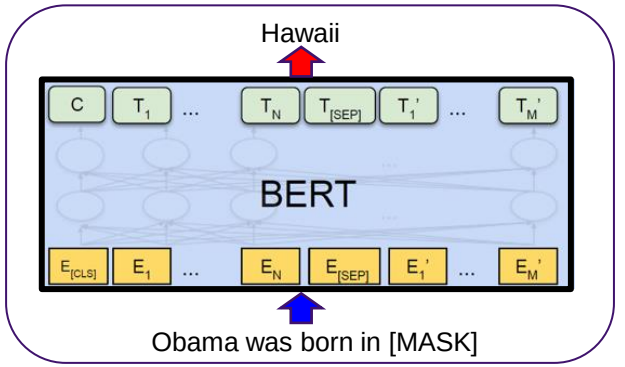
Step 1: pre-train (transformer-based) language model on large amounts of text

Step 2: fine-tune model on task
or: use prompting

Natural Language Processing
Current State of the Art







Step 1: pre-train (transformer-based) language model on large amounts of text

Step 2: fine-tune model on task
or: use prompting

Natural Language Processing
Current Limitations

- Pre-trained language models usually achieve high accuracy on benchmark datasets

Natural Language Processing
Current Limitations

- Pre-trained language models usually achieve high accuracy on benchmark datasets
- However, this does not mean that they are usable in real-world settings!

Natural Language Processing
Current Limitations

- Pre-trained language models usually achieve high accuracy on benchmark datasets
- However, this does not mean that they are usable in real-world settings!

- Problems (among others):
  - Focus on high-resource domains and languages
    - Not realistic: in reality we don't have labelled data for the tasks / domains / languages of interest
  - Black-box models
    - Explainability/interpretability required when deploying systems for real users

Natural Language Processing
Current Limitations

- Pre-trained language models usually achieve high accuracy on benchmark datasets
- However, this does not mean that they are usable in real-world settings!

- Problems (among others):
  - Focus on high-resource domains and languages
    - Not realistic: in reality we don't have labelled data for the tasks / domains / languages of interest
  - Black-box models
    - Explainability/interpretability required when deploying systems for real users

This talk:

Low-Resource NLP

Explainable NLP

Low-Resource NLP
Motivation

- Current research focuses on high-resource domains and languages
  - Current deep-learning models are data hungry
  - Not realistic: In reality we don’t have labelled data for the tasks / domains / languages of interest
  - Possible directions: [Hedderich & Lange et al., NAACL 2021]
    - Creation of additional data (e.g., data augmentation, weak supervision)
    - Make use of unsupervised data (e.g., pre-trained language models)
    - Reduction of need of supervision (e.g., transfer learning)

Low-Resource NLP
Motivation

- Current research focuses on high-resource domains and languages
  - Current deep-learning models are data hungry
  - Not realistic: In reality we don't have labelled data for the tasks / domains / languages of interest
  - Possible directions: [Hedderich & Lange et al., NAACL 2021]
    - Creation of additional data (e.g., data augmentation, weak supervision)
    - Make use of unsupervised data (e.g., pre-trained language models)
    - Reduction of need of supervision (e.g., transfer learning)

- Our approaches to cope with this challenge:
  - Meta-embeddings [Lange et al., EMNLP 2021]
  - Selection of Transfer Sources [Lange et al., EMNLP 2021]

Joint work with:

Lukas Lange, BCAI

Jannik Strötgen, BCAI

Dietrich Klakow, Saarland University

Low-Resource NLP
Motivation

- Current research focuses on high-resource domains and languages
  - Current deep-learning models are data hungry
  - Not realistic: In reality we don't have labelled data for the tasks / domains / languages of interest
  - Possible directions: [Hedderich & Lange et al., NAACL 2021]
    - Creation of additional data (e.g., data augmentation, weak supervision)
    - Make use of unsupervised data (e.g., pre-trained language models)
    - Reduction of need of supervision (e.g., transfer learning)

- Our approaches to cope with this challenge:
  - Meta-embeddings [Lange et al., EMNLP 2021]
  - Selection of Transfer Sources [Lange et al., EMNLP 2021] focus of this talk

Joint work with:

Lukas Lange, BCAI

Jannik Strötgen, BCAI

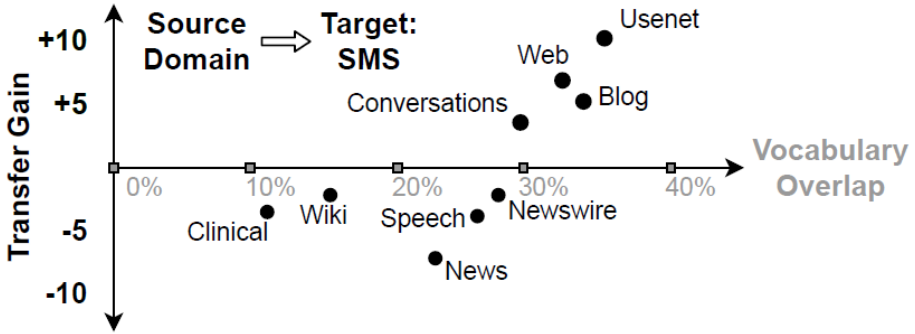
Dietrich Klakow, Saarland University

Low-Resource NLP

Transfer Learning: Selection of Transfer Sources

[Lange, Strötgen, Adel & Klakow, EMNLP 2021]

- Observation: Different performance gains can be expected for different transfer sources



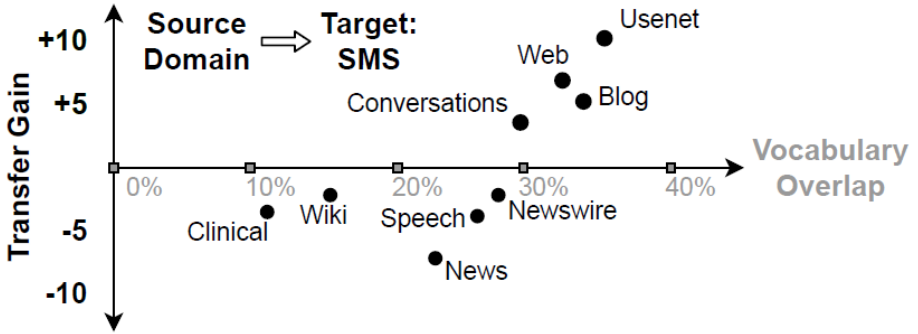
- For some transfer sources, the performance might actually drop (negative transfer gain)
- How to avoid negative transfer gains?
- How to select the set of most promising sources?

Low-Resource NLP

Transfer Learning: Selection of Transfer Sources

[Lange, Strötgen, Adel & Klakow, EMNLP 2021]

- Observation: Different performance gains can be expected for different transfer sources



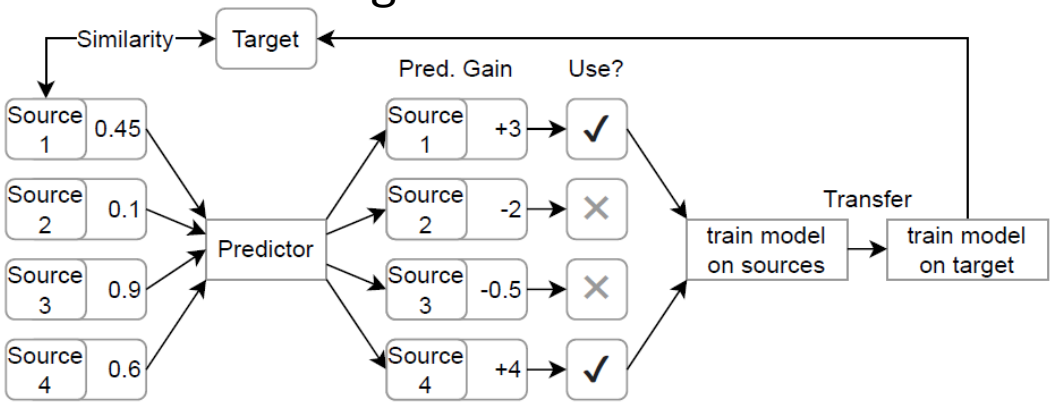
- For some transfer sources, the performance might actually drop (negative transfer gain)
- How to avoid negative transfer gains?
- How to select the set of most promising sources?

- Proposal: New similarity measure and new method for selecting set of sources:

Model similarity measure:

$$MoS(m_s, m_t) = |W - I|$$

$$\arg\min_W |W(f(m_s, t)) - f(m_t, t)|$$

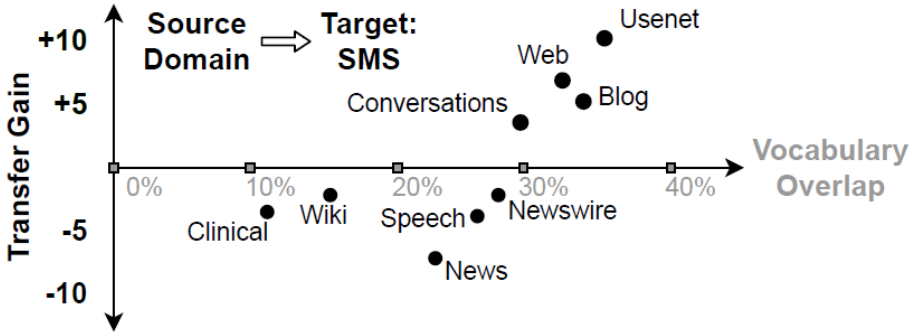


Low-Resource NLP

Transfer Learning: Selection of Transfer Sources

[Lange, Strötgen, Adel & Klakow, EMNLP 2021]

- Observation: Different performance gains can be expected for different transfer sources



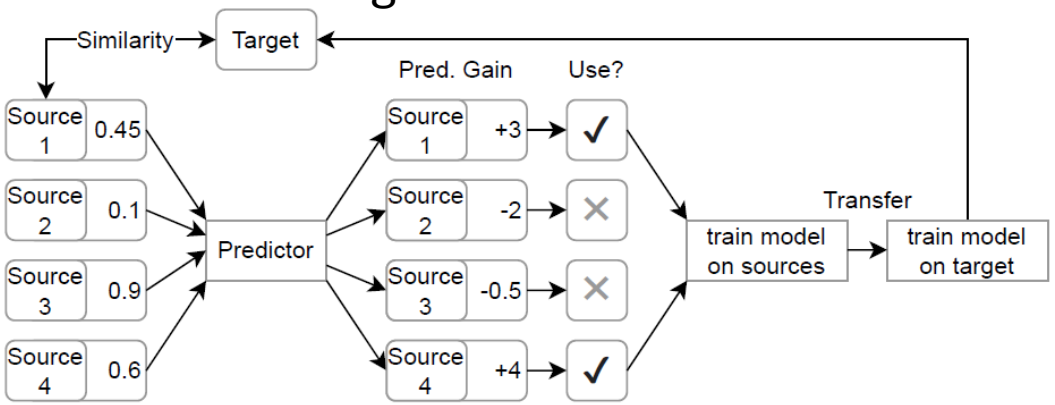
- For some transfer sources, the performance might actually drop (negative transfer gain)
- How to avoid negative transfer gains?
- How to select the set of most promising sources?

- Proposal: New similarity measure and new method for selecting set of sources:

Model similarity measure:

$$MoS(m_s, m_t) = |W - I|$$

$$\arg\min_W |W(f(m_s, t)) - f(m_t, t)|$$



- **Take-home message:** Transfer learning is not successful by default: Carefully select the sources!

Explainable NLP
Motivation

- Current research focuses on black-box models
  - Explainability/interpretability required when deploying systems for real users
  - Another observation: Humans might not interpret model outputs or explanations as expected

Explainable NLP
Motivation

- Current research focuses on black-box models
  - Explainability/interpretability required when deploying systems for real users
  - Another observation: Humans might not interpret model outputs or explanations as expected

- Our approaches to cope with this challenge:
  Models → Evaluation → Perception
  - Coupling of prediction and explanation [Schuff et al., EMNLP 2020]
  - Evaluation scores [Schuff et al., EMNLP 2020]
  - Human interpretation of attribution scores [Schuff et al., FAccT 2022]

Joint work with:

Hendrik Schuff, BCAI

Ngoc Thang Vu, University of Stuttgart

BOSCH

Explainable NLP
Motivation

- Current research focuses on black-box models
  - Explainability/interpretability required when deploying systems for real users
  - Another observation: Humans might not interpret model outputs or explanations as expected

- Our approaches to cope with this challenge:
  Models → Evaluation → Perception
  - Coupling of prediction and explanation [Schuff et al., EMNLP 2020]
  - Evaluation scores [Schuff et al., EMNLP 2020]
  - Human interpretation of attribution scores focus of this talk
    [Schuff et al., FAccT 2022]

Joint work with:

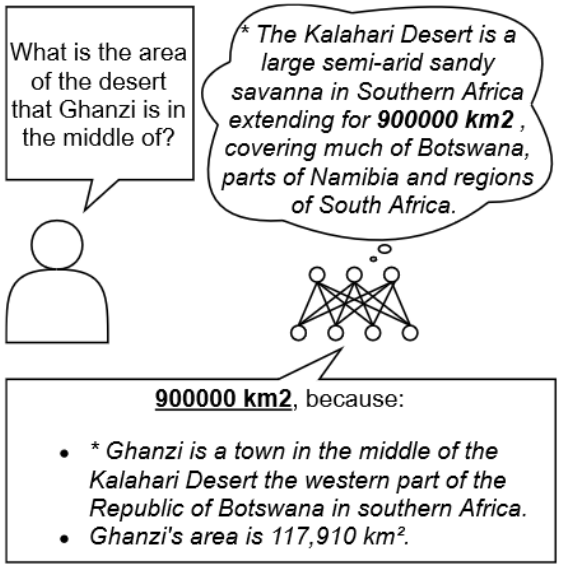
Hendrik Schuff, BCAI

Ngoc Thang Vu, University of Stuttgart

BOSCH

Explainable NLP

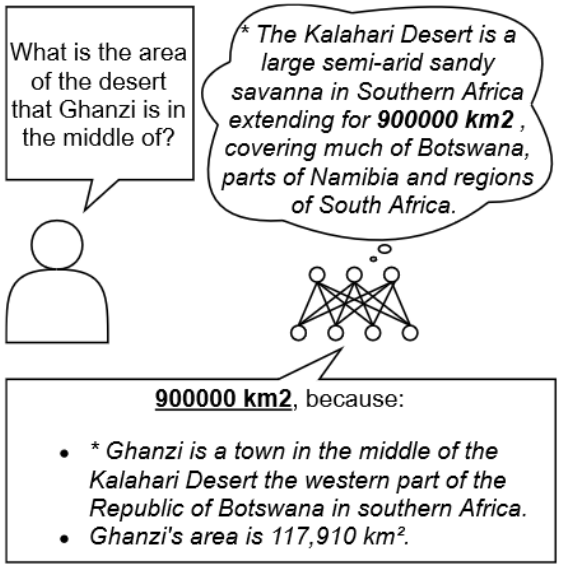
Models: Coupling of Predictions and Explanations [Schuff, Adel & Vu, EMNLP 2020]



- Observation: Task prediction and explanation are not necessarily coupled in state-of-the-art models
- Current evaluation scores do not take this into account

Explainable NLP

Models: Coupling of Predictions and Explanations [Schuff, Adel & Vu, EMNLP 2020]



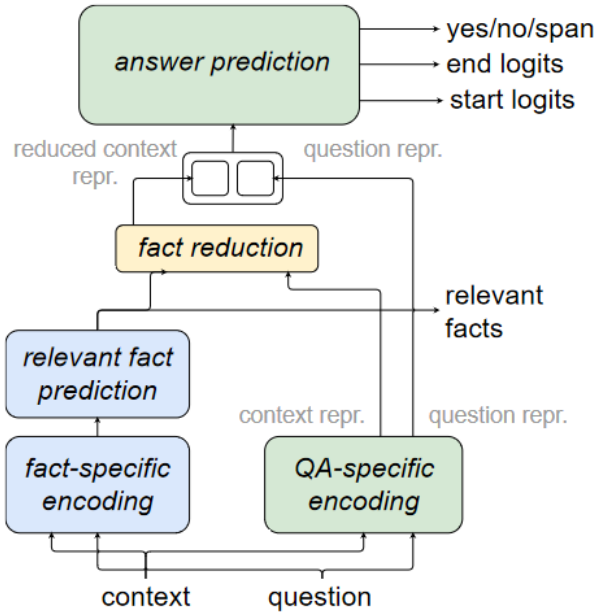
- Observation: Task prediction and explanation are not necessarily coupled in state-of-the-art models
- Current evaluation scores do not take this into account

- Proposal: Novel model architecture and loss function for ensuring coupling

Regulization term for loss:

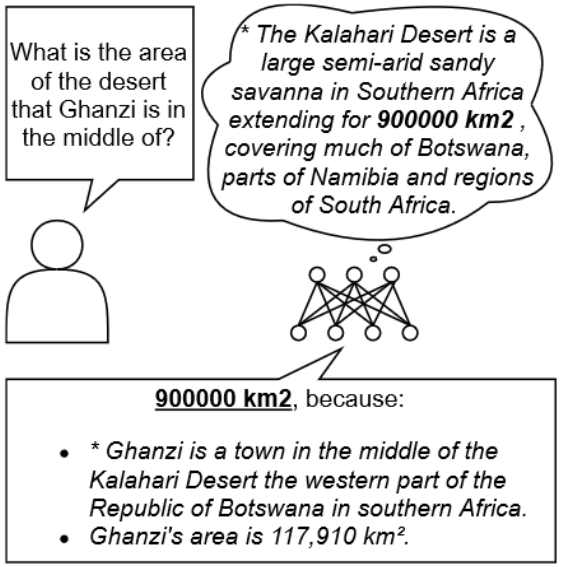
$$J_{\text{reg}} = \underbrace{p_a \cdot (\overbrace{p_e \cdot 0}^{\text{GT expl.}} + \overbrace{(1-p_e) \cdot c_1}^{\text{non-GT expl.}})}_{\text{correct answer}} + \underbrace{(1-p_a) \cdot (\overbrace{p_e \cdot c_2}^{\text{GT expl.}} + \overbrace{(1-p_e) \cdot c_3}^{\text{non-GT expl.}})}_{\text{wrong answer}}$$

Select-and-forget architecture:



Explainable NLP

Models: Coupling of Predictions and Explanations [Schuff, Adel & Vu, EMNLP 2020]

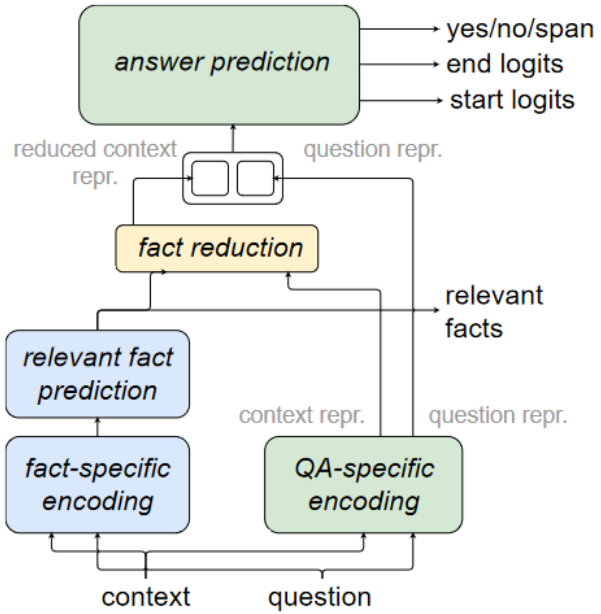


- Observation: Task prediction and explanation are not necessarily coupled in state-of-the-art models
- Current evaluation scores do not take this into account

Regulization term for loss:

$$J_{\text{reg}} = \underbrace{p_a \cdot (\overbrace{p_e \cdot 0}^{\text{GT expl.}} + \overbrace{(1-p_e) \cdot c_1}^{\text{non-GT expl.}})}_{\text{correct answer}} + \underbrace{(1-p_a) \cdot (\overbrace{p_e \cdot c_2}^{\text{GT expl.}} + \overbrace{(1-p_e) \cdot c_3}^{\text{non-GT expl.}})}_{\text{wrong answer}}$$

Select-and-forget architecture:

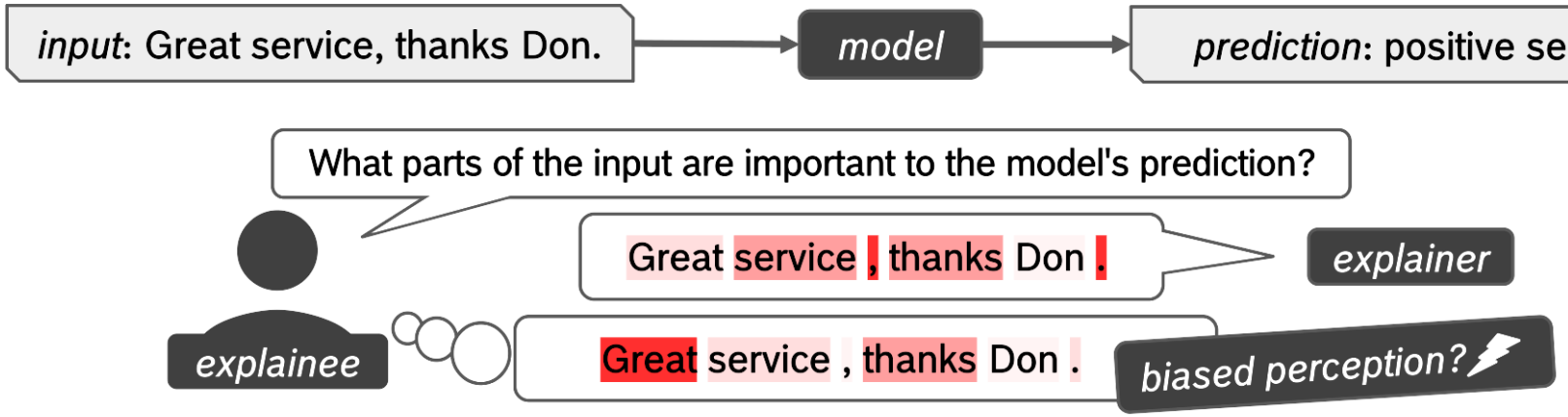


- Proposal: Novel model architecture and loss function for ensuring coupling
- **Take-home message:** Make sure that the explanations are coupled with the predictions of the model => otherwise they cannot be helpful for a user

Explainable NLP

Perception: Human Interpretation of Attribution Scores

[Schuff, Jacovi, Adel, Goldberg & Vu, FAccT 2022]

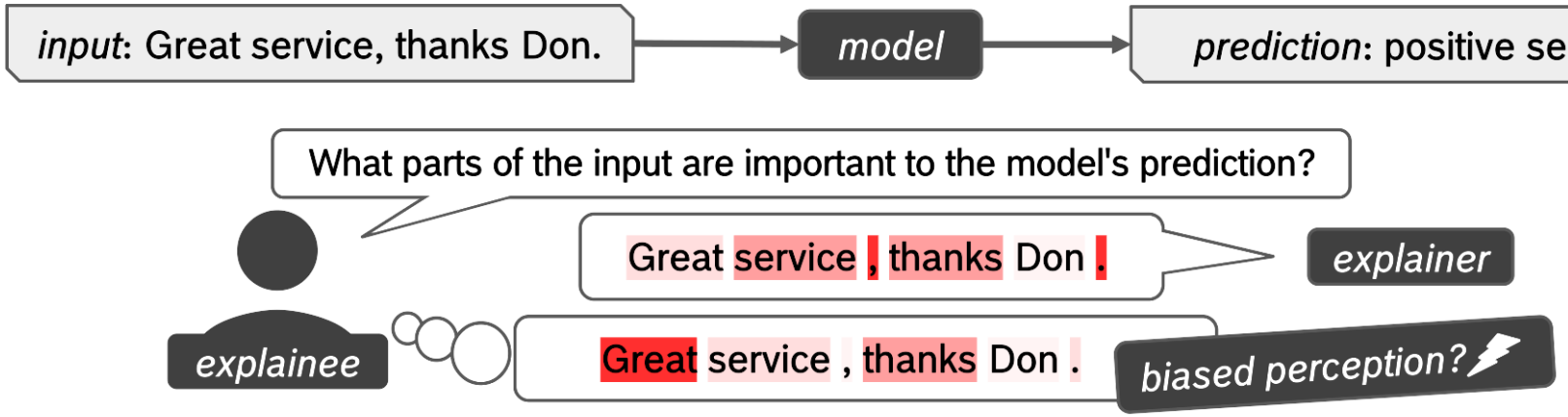


- Study: Which factors influence human interpretation of attribution scores?
  - Crowdsourcing user study (135 participants, >20k importance ratings, different languages / tasks / scores)
  - GAMM (Generalized Additive Mixed Model) for statistical analysis
  - Result: many biasing factors, e.g., word length, word position, capitalization, sentence length, saliency rank, …

Explainable NLP

Perception: Human Interpretation of Attribution Scores

[Schuff, Jacovi, Adel, Goldberg & Vu, FAccT 2022]

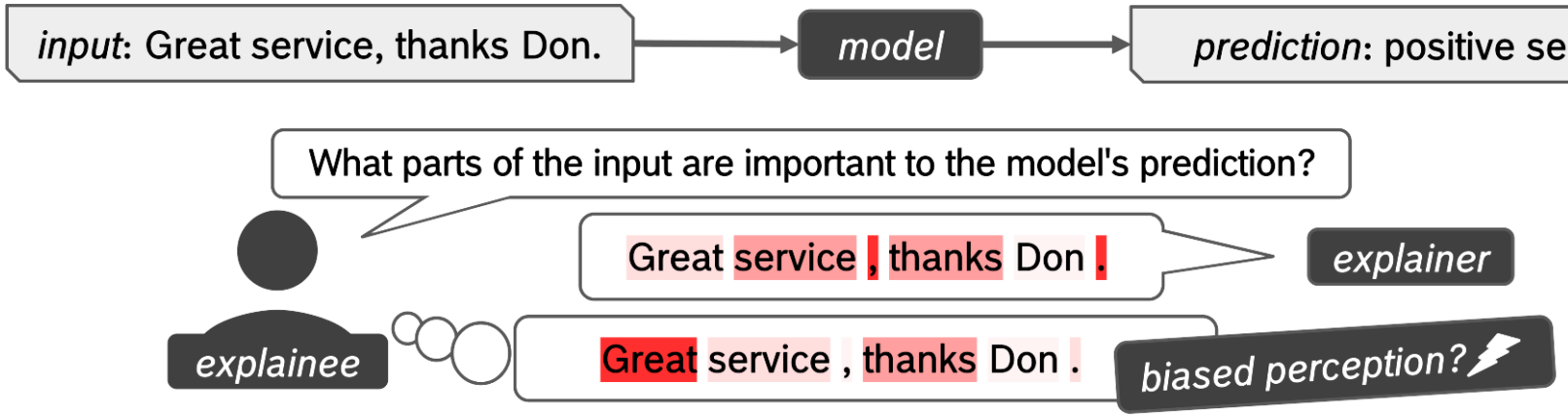


- Study: Which factors influence human interpretation of attribution scores?
  - Crowdsourcing user study (135 participants, >20k importance ratings, different languages / tasks / scores)
  - GAMM (Generalized Additive Mixed Model) for statistical analysis
  - Result: many biasing factors, e.g., word length, word position, capitalization, sentence length, saliency rank, ...
- Proposal to mitigate bias with using bar charts or adjusting salience scores

Explainable NLP

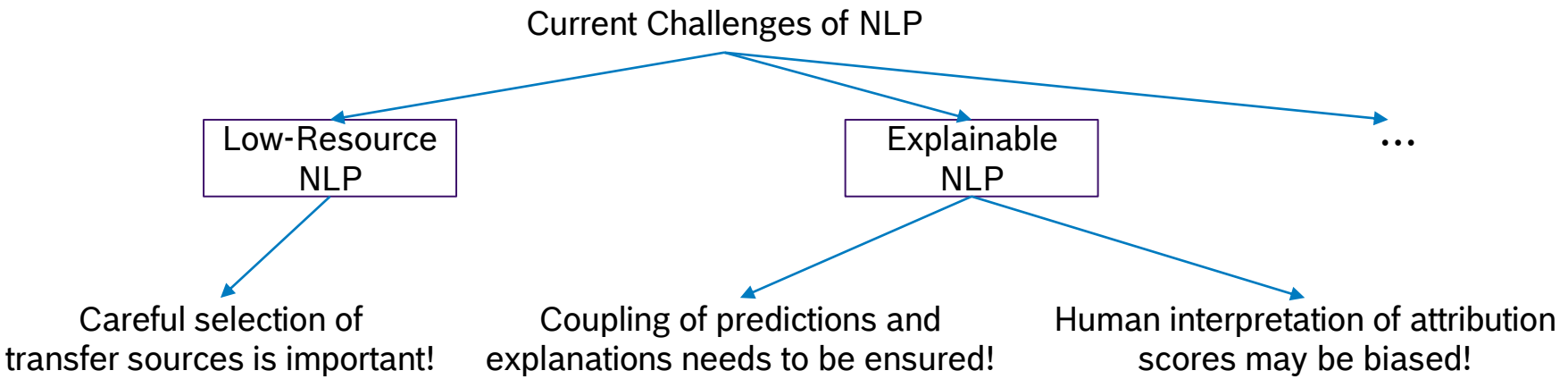
Perception: Human Interpretation of Attribution Scores

[Schuff, Jacovi, Adel, Goldberg & Vu, FAccT 2022]

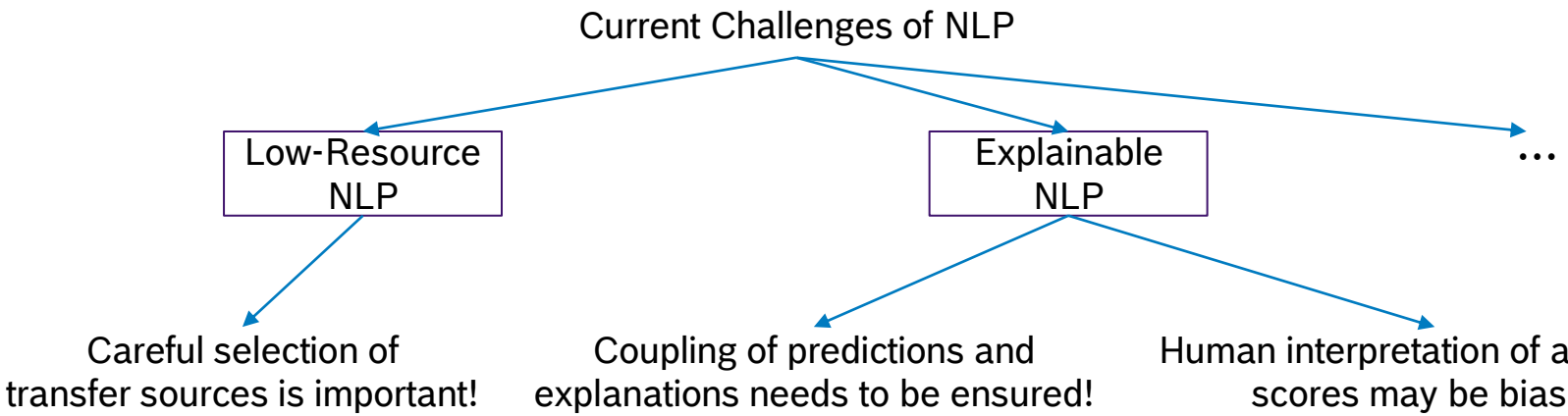


- Study: Which factors influence human interpretation of attribution scores?
  - Crowdsourcing user study (135 participants, >20k importance ratings, different languages / tasks / scores)
  - GAMM (Generalized Additive Mixed Model) for statistical analysis
  - Result: many biasing factors, e.g., word length, word position, capitalization, sentence length, saliency rank, …
- Proposal to mitigate bias with using bar charts or adjusting salience scores
- **Take-home message**: Human perception of saliency explanations might be biased: Take this into account to ensure that explanations are actually helpful for users!

Conclusion
Summary and Outlook

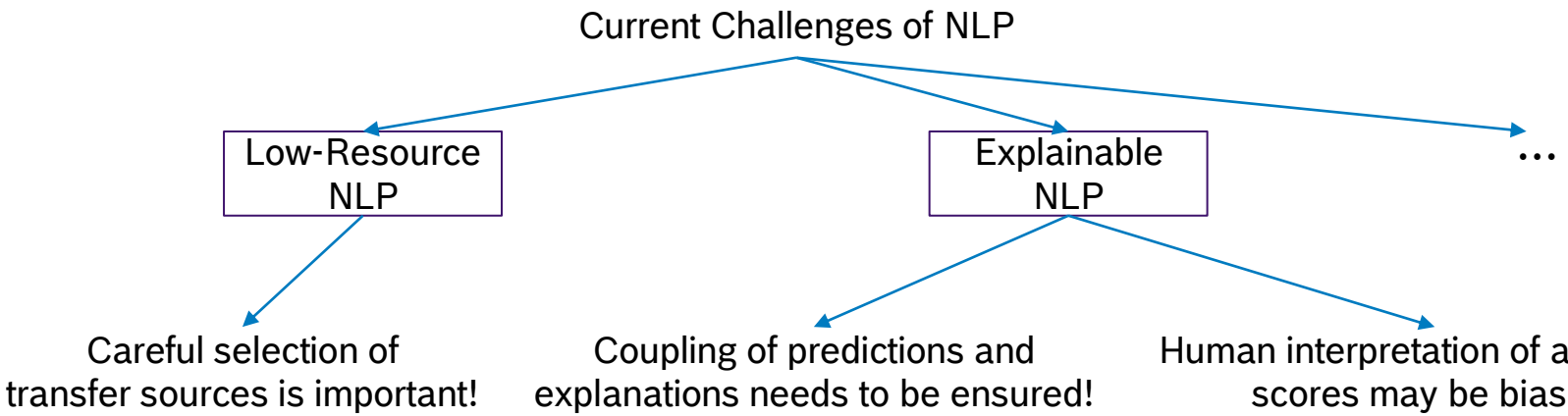


Conclusion
Summary and Outlook



- This talk presented some steps towards natural language processing in real-world settings

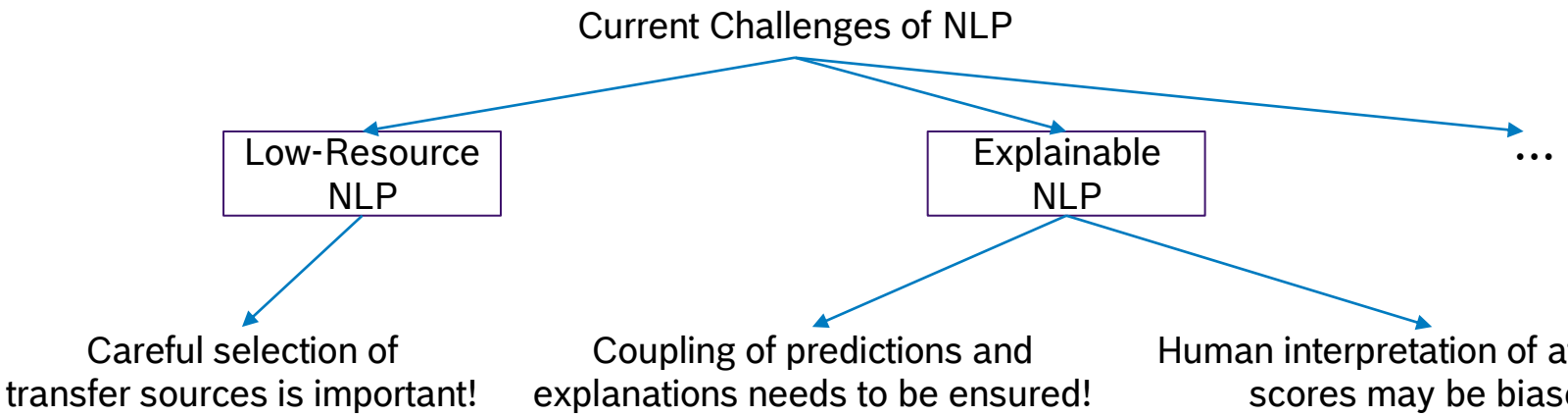
Conclusion
Summary and Outlook



- This talk presented some steps towards natural language processing in real-world settings
- However, we still have a long road ahead of us
- Exemplary further directions:
  - How to effectively augment training data for NLP? Low-Resource NLP
  - How to analyze pre-trained language models and make best use of their “knowledge”? Explainable NLP

Conclusion
Summary and Outlook

Thank you for your attention!



- This talk presented some steps towards natural language processing in real-world settings
- However, we still have a long road ahead of us
- Exemplary further directions:
  - How to effectively augment training data for NLP? Low-Resource NLP
  - How to analyze pre-trained language models and make best use of their “knowledge”? Explainable NLP