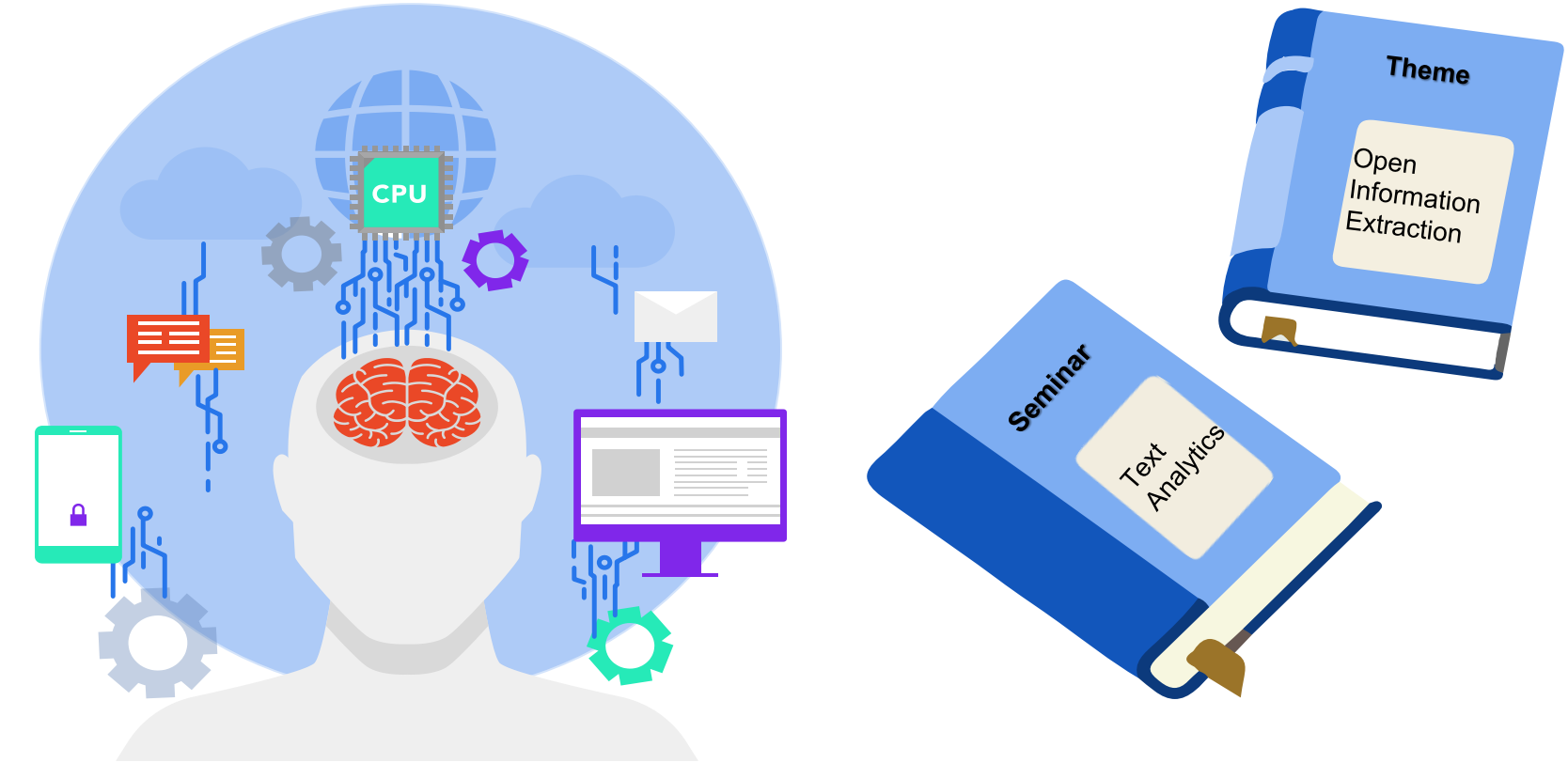
DocOIE : A Document-level Context-Aware Dataset for OpenIE

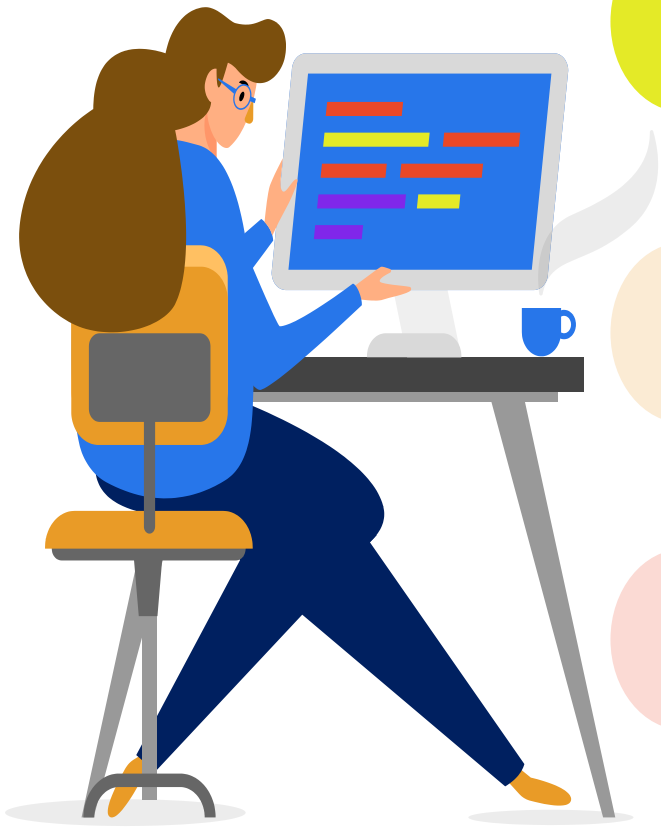
Kuicai Dong, Yilin Zhao, Aixin Sun, Jung-Jae Kim, Xiaoli Li

Presented by : Amira Chebbi

TECHNISCHE UNIVERSITÄT DARMSTADT



Overview



01 Introduction

02 Background and related work

03 Methodology

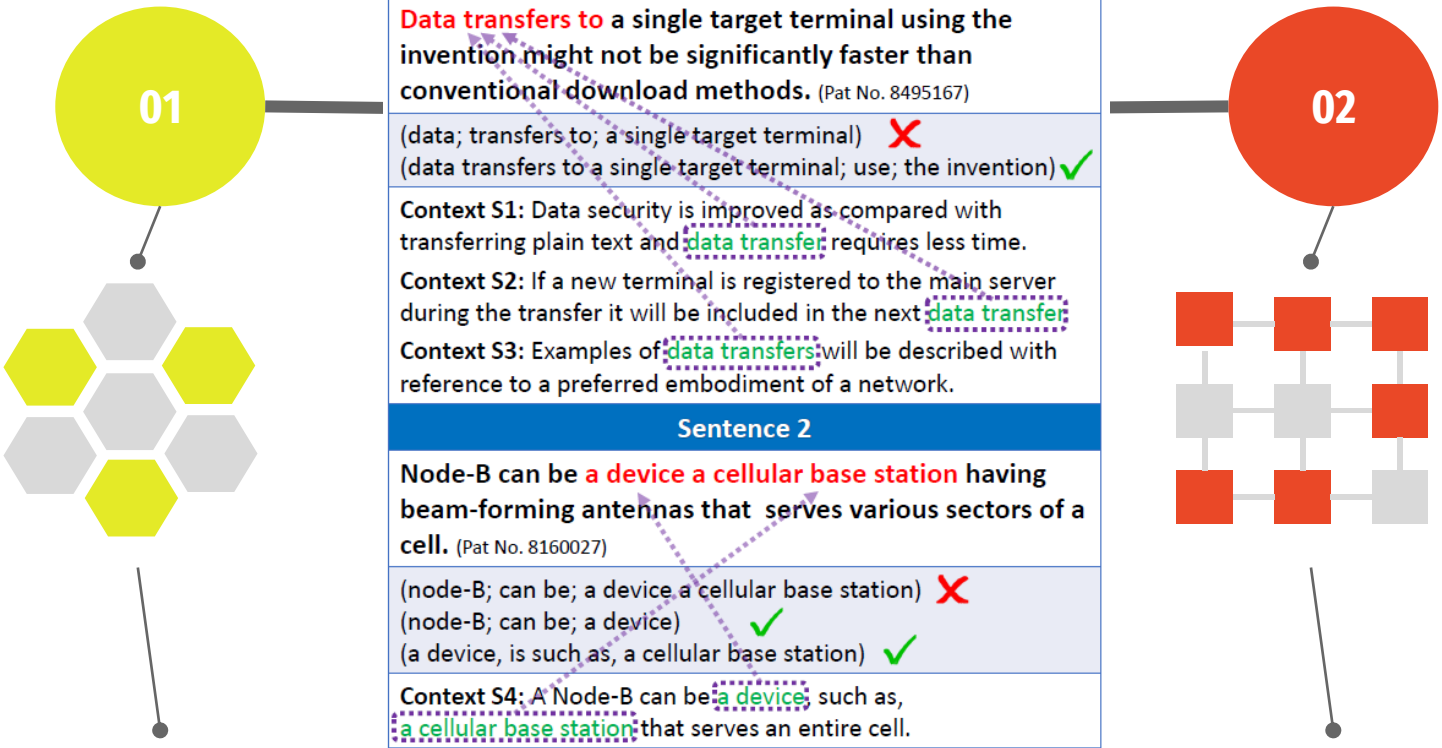
04 Results

05 Analysis

06 Bibliography

Introduction - **Problematic**

DocOIE : A Document-level Context-Aware Dataset for OpenIE



Part-of-speech Ambiguity

Is it a Verb or a Noun ?

Syntactic Ambiguity

What is the relationship between this two words ?

Introduction - **Proposed Solution**

DocOIE : A Document-level Context-Aware Dataset for OpenIE

TECHNISCHE UNIVERSITÄT DARMSTADT

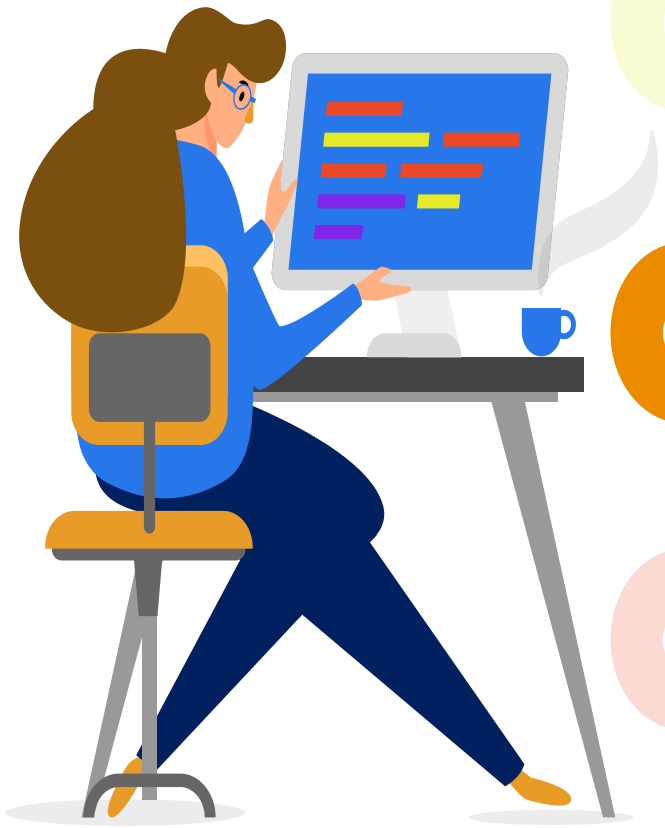
DocOIE proposed solutions

01 **New Task in OpenIE**
Extract relational tuples with document-level context

02 **DocOIE**
Expert-annotated dataset

03 **DocIE**
Neural OpenIE system

Overview



01 Introduction

02 Background and related work

03 Methodology

04 Results

05 Analysis

06 Bibliography

Related work - **OpenIE Datasets**

TECHNISCHE UNIVERSITÄT DARMSTADT

OpenIE Datasets	
OIE2016 (Gabriel Stanovsky and Ido Dagan. 2016)	first large-scale dataset constructed for OpenIE tasks and comes with a standard scoring framework.the gold tuples are automatically generated according to human crafted rules.
Wire57(Lechelle et al., 2019)	manually annotates 57 sentences + scoring improvement
CaRB (Bhardwaj et al. (2019))	50 expert-annotated sentences + a sophisticated scoring framework

Related work - OpenIE Datasets - **Problems**

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset's size problem

The number of the annotated sentences ramains small

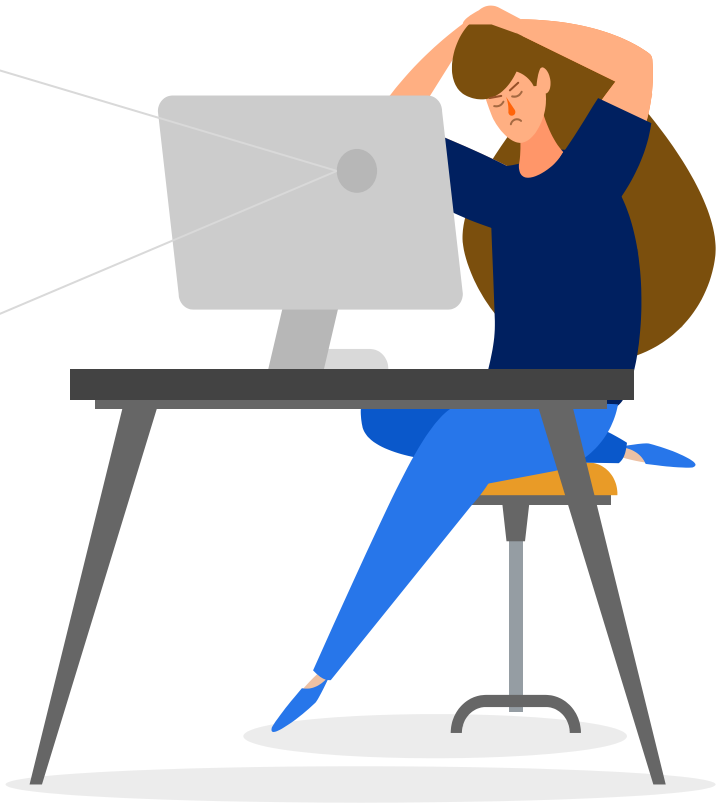
01

No referring to contextual Information

The Annotation is elaborated on a sentence-level and not a document-level

02

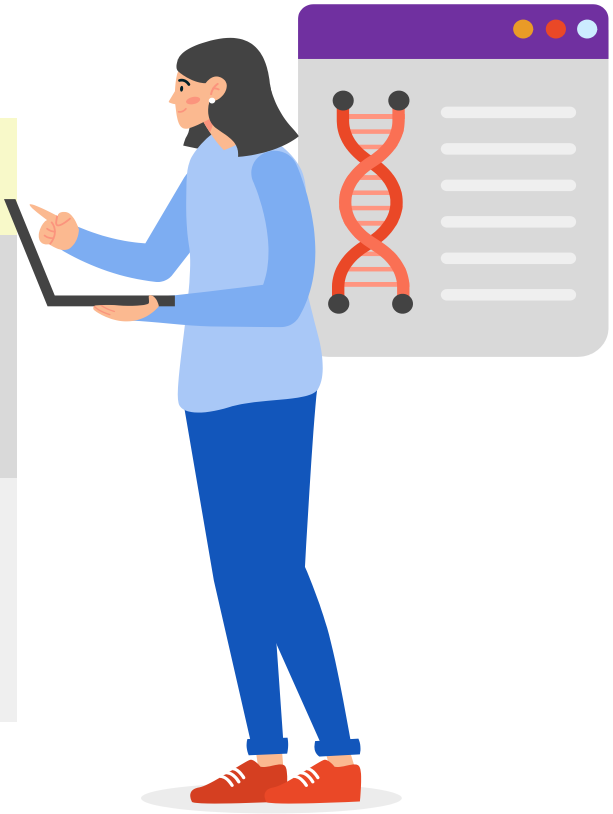
OpenIE Datasets Problems



Related works - **OpenIE Models**

TECHNISCHE UNIVERSITÄT DARMSTADT

OpenIE Models	
TextRunner (Yates et al., 2007)	the first OpenIE system,extract relational tuples based on handcrafted rules or statistical methods
neural OpenIE systems (Cui et al., 2018; Zhan and Zhao, 2020; Kolluru et al., 2020a,b)	They extract tuples in an end-toend manner, not requiring prior syntactic or semantic analysis.



Related works - OpenIE Models - **Problems**

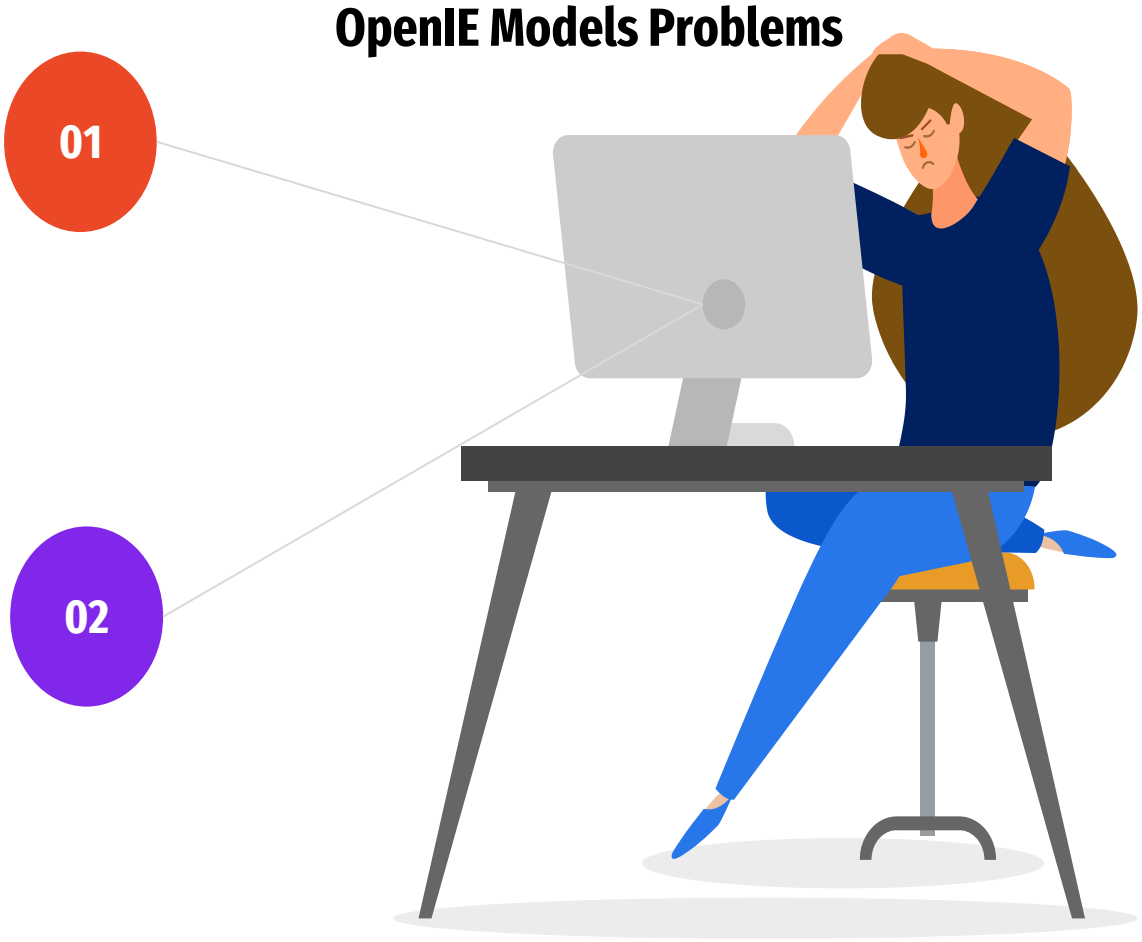
TECHNISCHE UNIVERSITÄT DARMSTADT

Error accumulation for traditional models

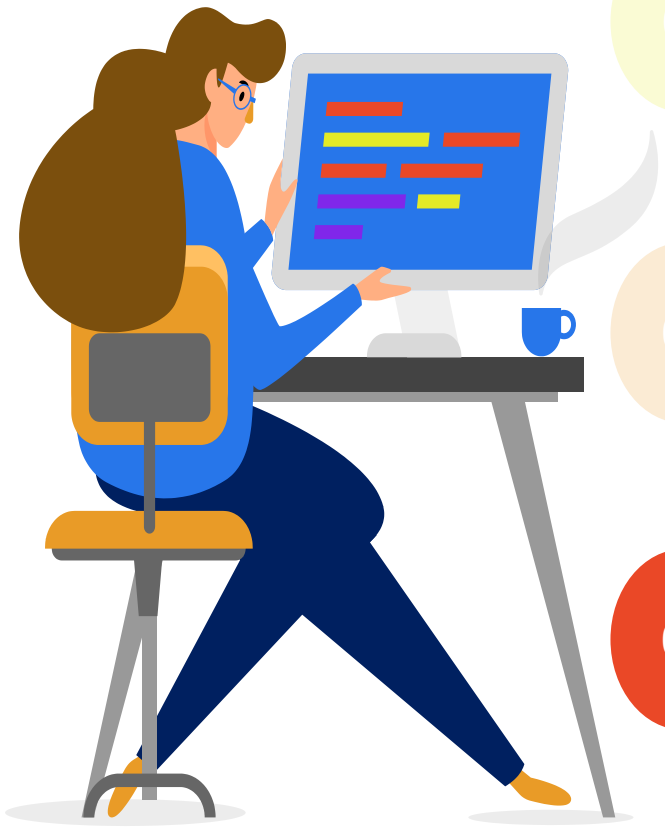
Because they rely on prior syntactic or semantic analysis

No consideration of the document-level context

The tuple extraction is based only on the sentence-level



Overview



01 Introduction

02 Background and related work

03 Methodology

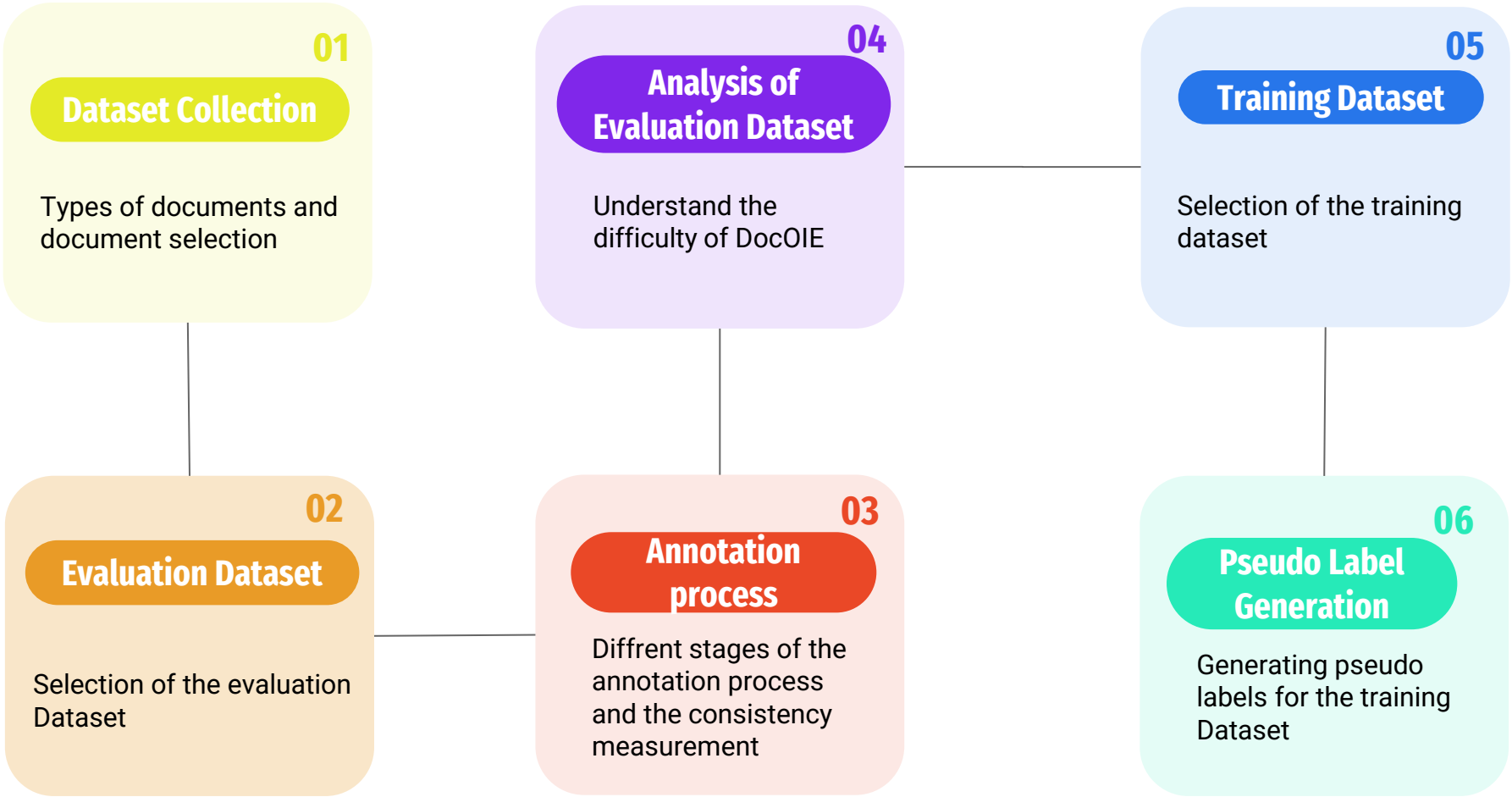
04 Results

05 Analysis

06 Bibliography

Methodology – **DocOIE Dataset**

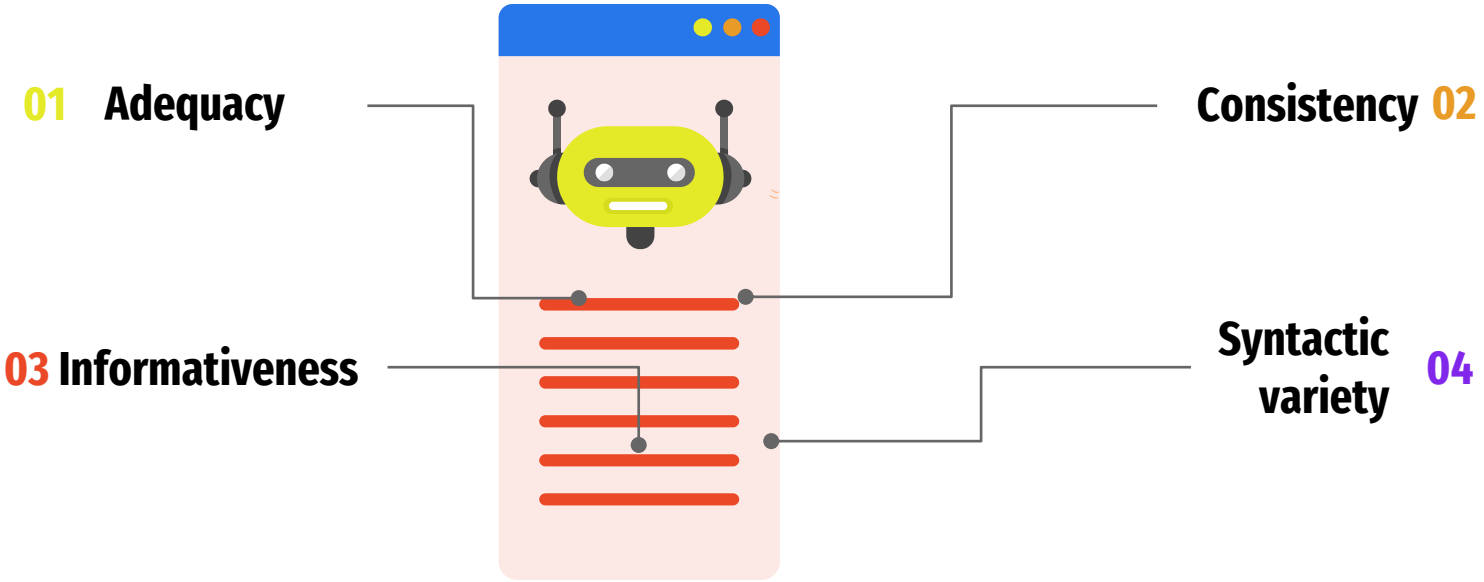
TECHNISCHE UNIVERSITÄT DARMSTADT



Methodology –DocOIE Dataset

1- Dataset Collection

- Document types : News and Wikipedia articles
- Document selection criterias :



- Keyword selection criterias : Magnitude and Diversity
- Document cleaning

Methodology –DocOIE Dataset
2-Evaluation Dataset

TECHNISCHE UNIVERSITÄT DARMSTADT

- Randomly select 80 documents (40 each domain)
- From each document , select randomly 10 sentences
- Result : 800 expert-annotated sentences from the DocOIE evaluation Dataset

Methodology –DocOIE Dataset

3- Annotation process

TECHNISCHE UNIVERSITÄT DARMSTADT

Annotation process and consistency measurement

Stage 1

1- Annotation

The two annotators practiced annotations independently on 100 sentences

2- Cross validation

Crossvalidation of the results , discussion and then update

Stage 2

1- Annotation

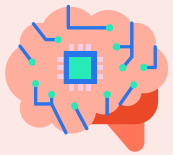
The two experts annotated independently other 100 sentences

2- Consistency Measurement

Evaluation used at tuple-level using the CaRB scorer

Stage 3

Based on the high-level annotation consistency , each expert annotate independently 300 sentences.



Methodology –DocOIE Dataset

4- Analysis of evaluation dataset

TECHNISCHE UNIVERSITÄT DARMSTADT

- Used to understand the difficulty of DocIE , similar to (Gashteovski et al., 2019).

Sentence-level Analysis

Evaluate sentence complexity by the number of :

- Conjunction words
- Terminology mention
- Dependent clause

and

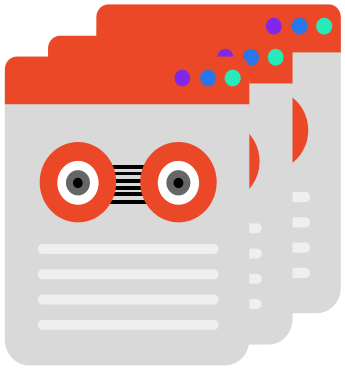
Tuple-level Analysis

The analysis of a tuple is based on three points :

- Negative polarity
- Possibility
- Under-specificity

Methodology –DocOIE Dataset

5- Trainig Dataset



- Select 2400 documents randomly (1200 each Domain)

- The 1200 documents in each domain contain around 120.000 sentences (sufficient for the openIE model)

Methodology –DocOIE Dataset

6- Pseudo Label generation

TECHNISCHE UNIVERSITÄT DARMSTADT

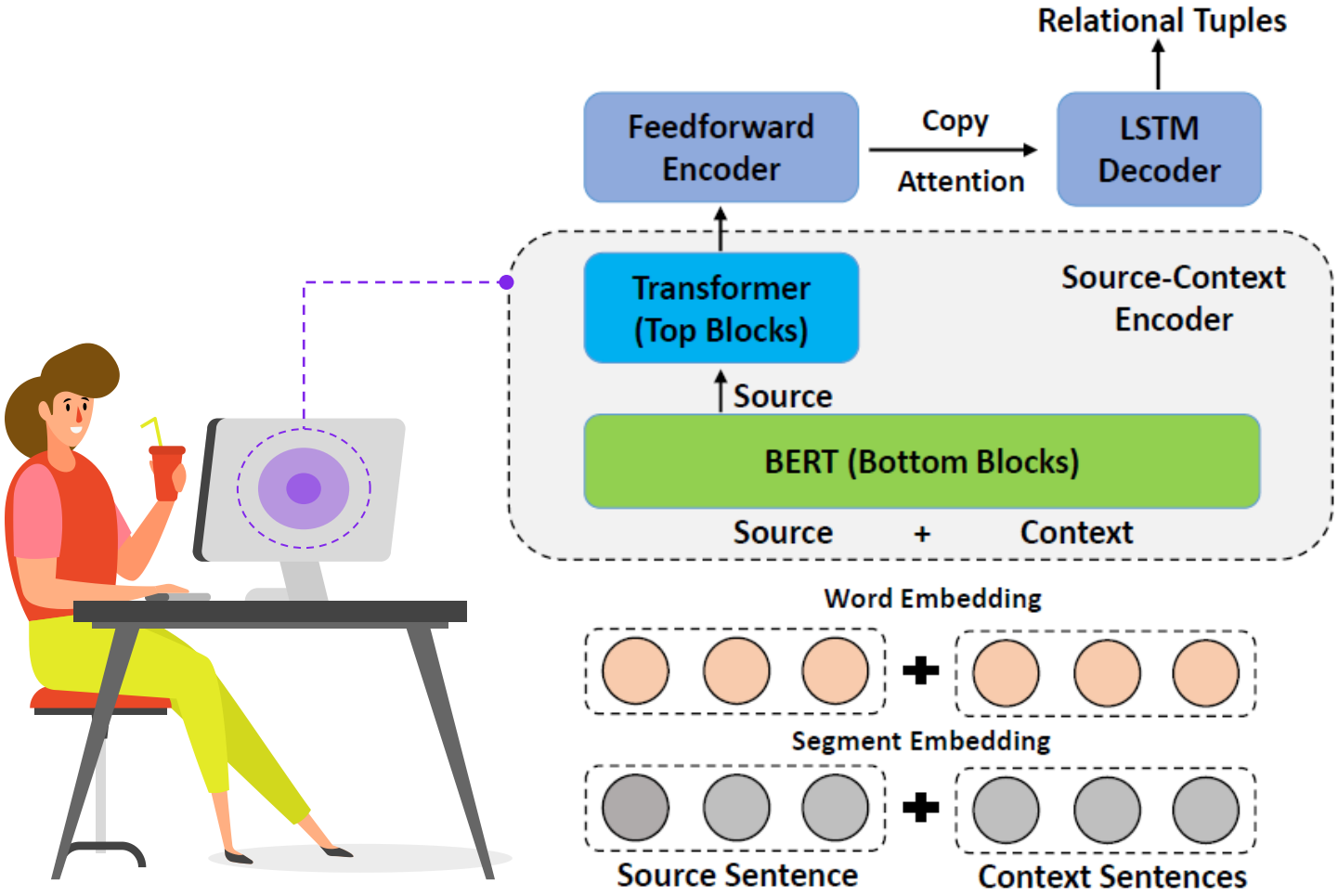
- Generation of the pseudo labels by bootstrapping with traditional OpenIE models(Kolluru et al.,2020b; Cui et al., 2018; Zhao et al., 2020)

- But First : Evaluation of the performance of the traditional OpenIE models on the evaluation Dataset using CaRB scorer(to guarantee better quality of pseudo labels)

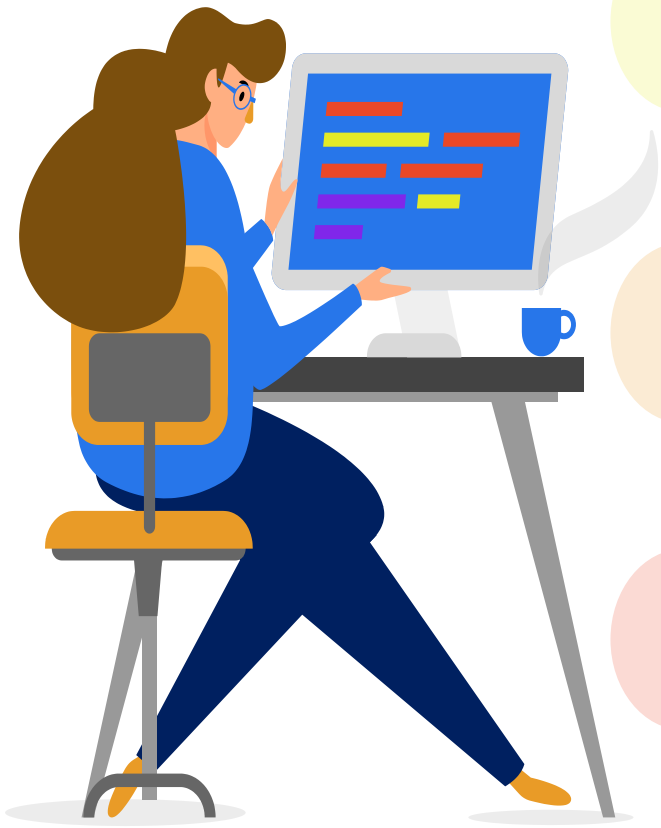
- The evaluated models were : Reverb (Fader et al., 2011), Clausie (Corro and Gemulla, 2013), Stanford OpenIE (Angeli et al.,2015), OpenIE4 (Mausam, 2016) , OpenIE5 , Rev+Oie4 and Oie4+ Rev

- Results shows that : both Reverb and OpenIE4 are the best performing individual models and their combinations lead to the best and second best F1 scores in both domains.

Methodology –DocIE Model



Overview



01 Introduction

02 Background and related work

03 Methodology

04 Results

05 Analysis

06 Bibliography

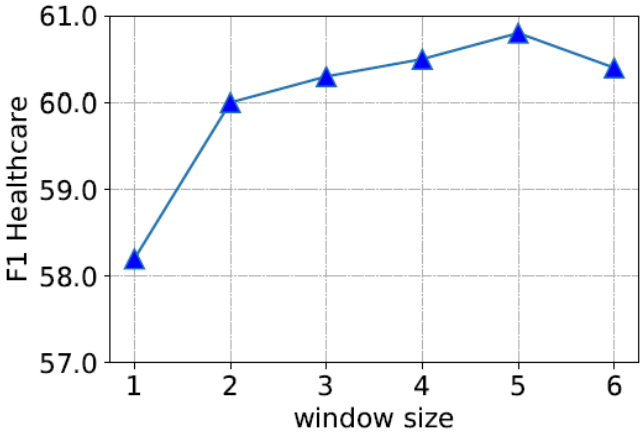
Results - DocIE Against sentence-level OpenIE systems

TECHNISCHE UNIVERSITÄT DARMSTADT

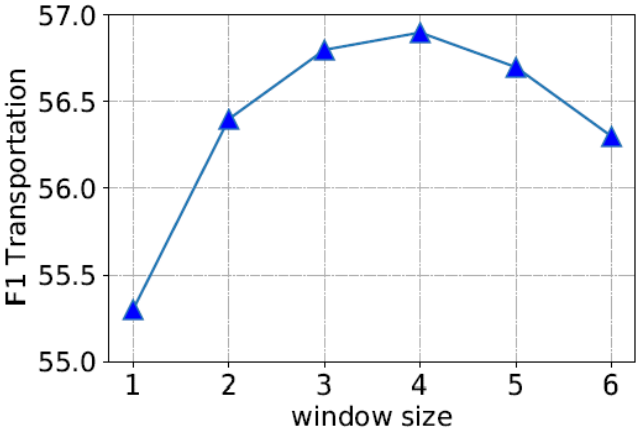
System	Healthcare				Transportation			
	AUC	Prec	Rec	F1	AUC	Prec	Rec	F1
Rev+Oie4	36.8	75.8	47.7	58.6	31.0	**74.2**	42.4	54.0
Oie4+Rev	35.8	59.6	**55.3**	57.4	30.1	53.4	52.7	53.0
CopyAttention+BERT	46.8	77.9	48.6	59.8	38.3	55.3	56.9	56.1
IMOJIE	39.7	**80.1**	46.4	58.7	35.8	63.5	49.2	55.5
DocIE w/o transformer	47.1	76.2	49.9	60.3	**38.5**	55.8	57.0	56.4
DocIE w transformer	**47.4**	74.4	51.3	**60.8**	**38.5**	56.0	**57.5**	**56.9**

- DocIE with transformer achieves the best AUC and F1 in both domains .
- DocIE without transformer is the second best performer and outperforms all the sentence-level systems .

Results - **Impact of the context Window size**



(a) Healthcare



(b) Transportation

The optimal windowsize for the healthcare is 5

The optimal windowsize for the transportation is 4

8-10 sentences provide sufficient context

and

Large window size may introduce noise

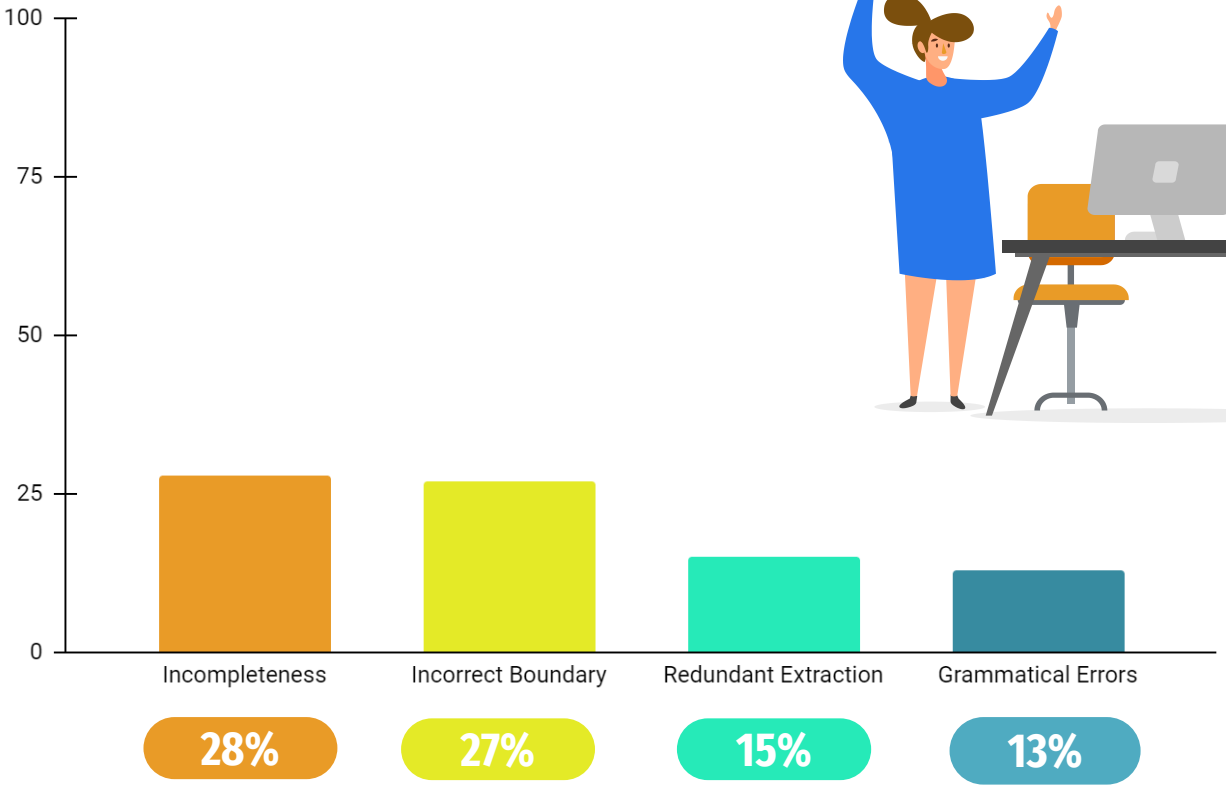
Results - Error Analysis

01 Incompleteness
Fails to cover at least one key phrase in either arguments or relations

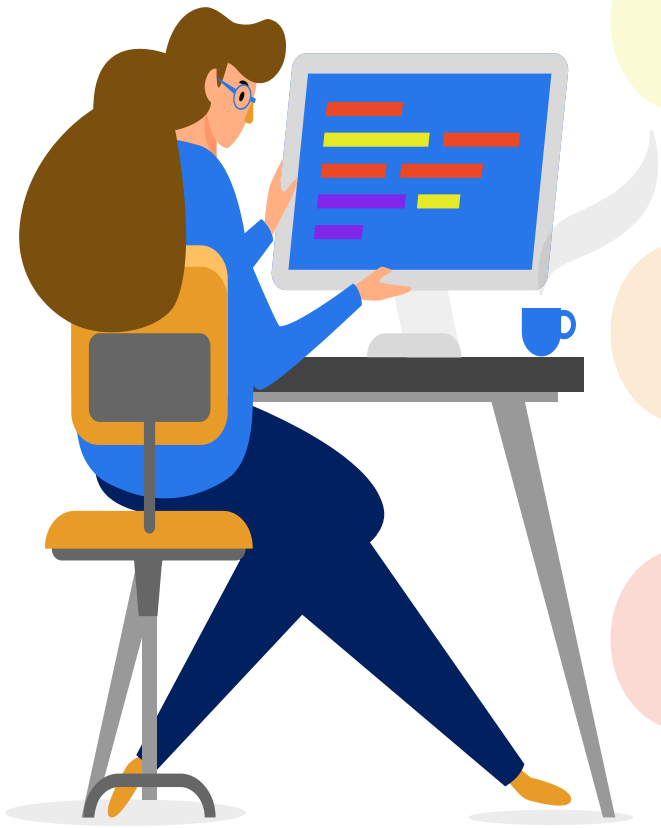
02 Incorrect Boundary
Misinterpretaion of the syntactic meaning of the sentence , leading to incorrect Boundary

03 Redundant extraction
The same relational fact is extracted multiple times

04 Grammatical Errors
Extracions are not grammatically correct



Overview



01 Introduction

02 Background and related work

03 Methodology

04 Results

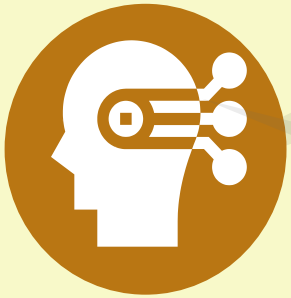
05 Analysis

06 Bibliography

Analysis – **Contributions**

TECHNISCHE UNIVERSITÄT DARMSTADT

Contribution-facet



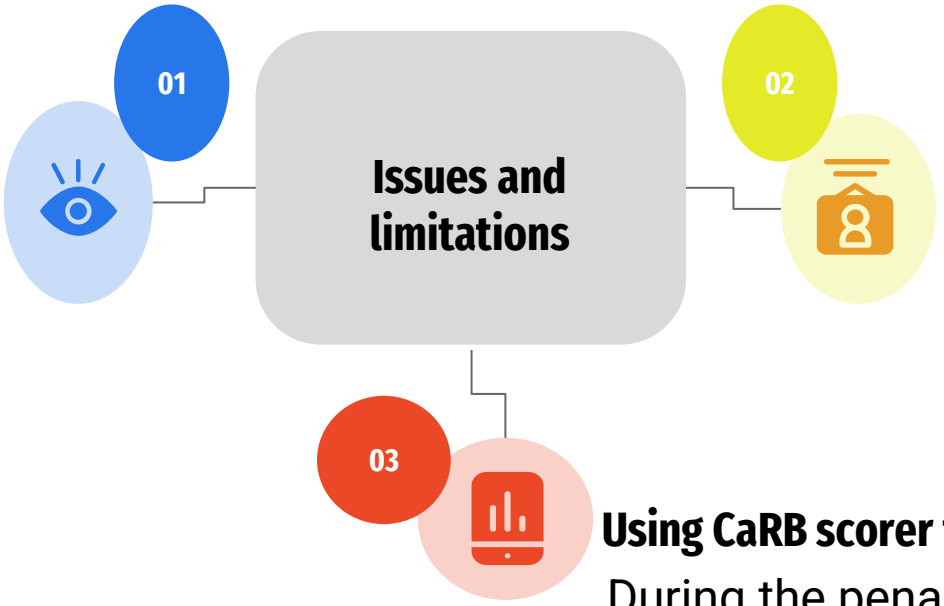
Positive outcomes

Better correct understanding of various topics First document-level context-aware OpenIE dataset and model

Relation to public policy

better analysis and helps in writing the adequate public policy and taking the right measurement in many areas.

Analysis – **Possible Issues and limitations**



No Generalization

DocIE Dataset is restricted to specific type of documents

Domains restriction

Foucus only on two domains : Transportion and Health-care

Using CaRB scorer for Evaluation

During the penalization CaRB scorer does not take the contextual sentences into consideration

Analysis – **Future Work**

TECHNISCHE UNIVERSITÄT DARMSTADT

01 **Model Improvement**

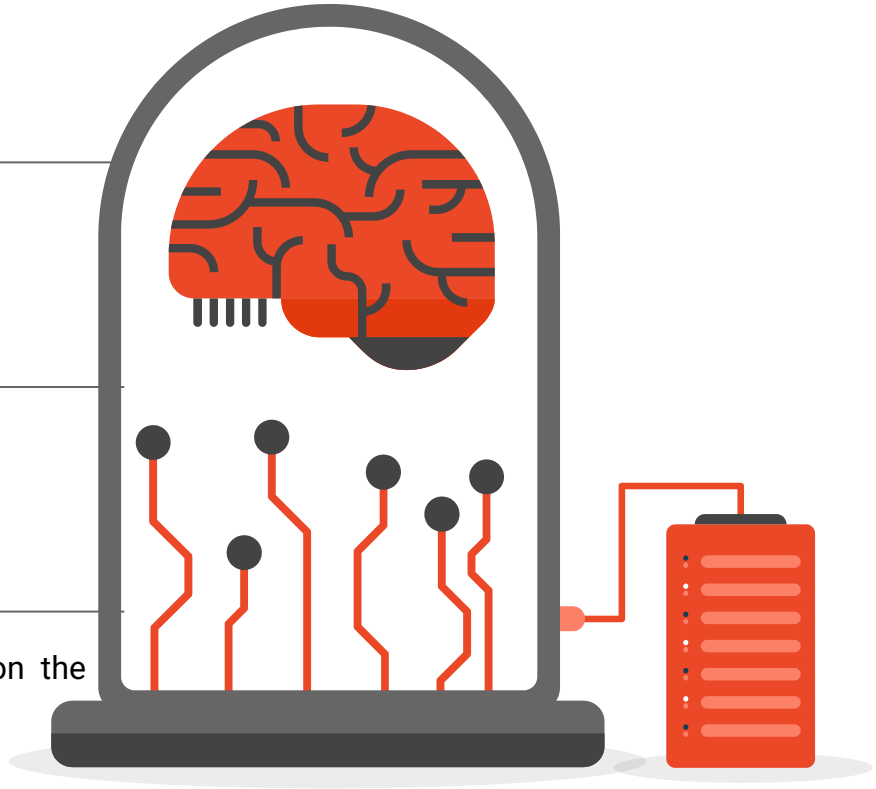
Research on more effective context-aware models

02 **Pseudo-Labels**

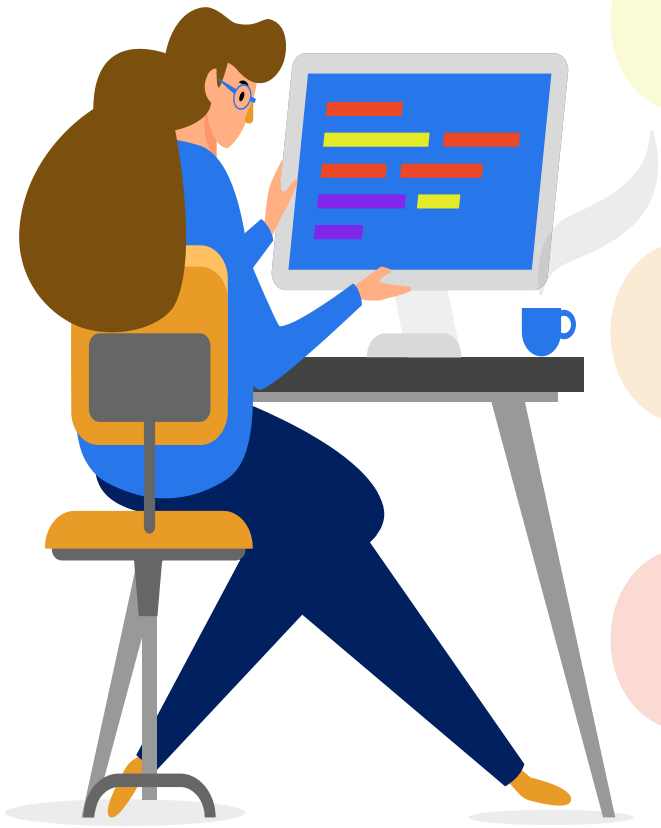
Invistigate the possiblity of not relying on pseudo labels

03 **Adapted Scorer**

Trying to develop a new scorer that takes into consideration the contextual informations



Overview



01 Introduction

02 Background and related work

03 Methodology

04 Results

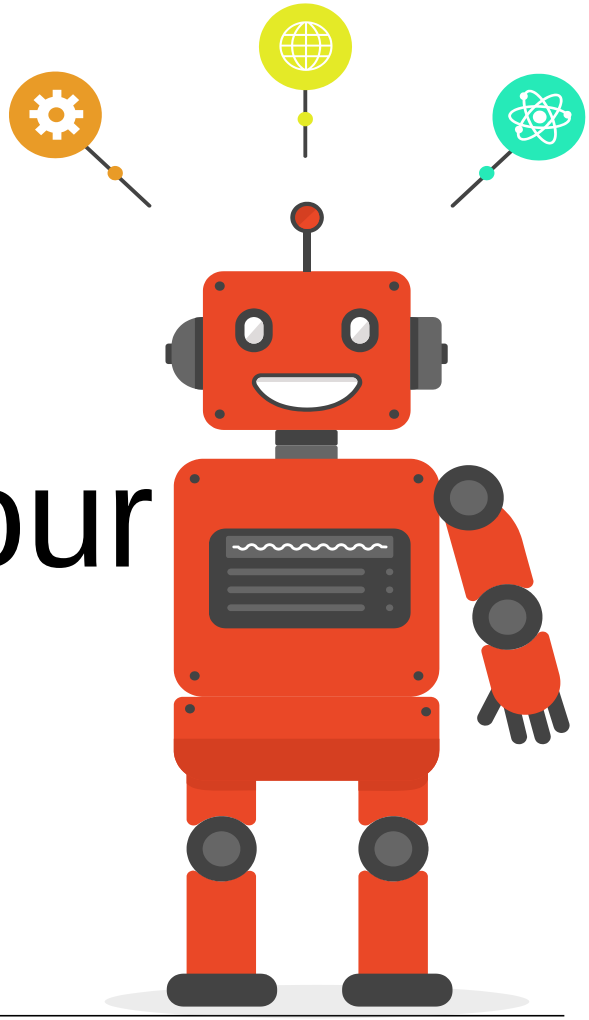
05 Analysis

06 Bibliography

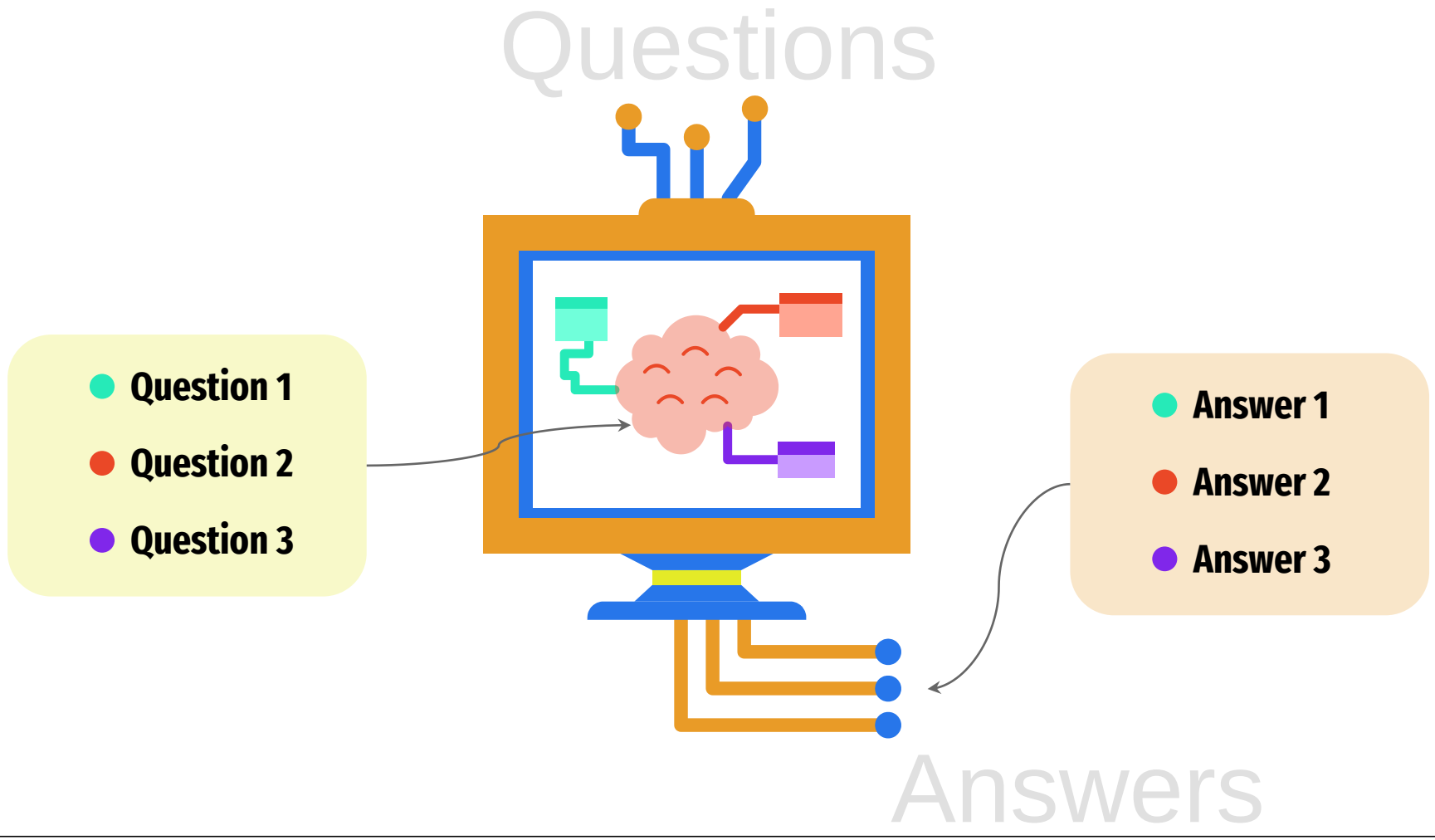
Refrences

- DocOIE: A Document-level Context-Aware Dataset for OpenIE
- Creating a Large Benchmark for Open Information Extraction
- WiRe57 : A Fine-Grained Benchmark for Open Information Extraction
- CaRB: A Crowdsourced Benchmark for Open IE
- TextRunner: Open Information Extraction on the Web
- Span Model for Open Information Extraction on Accurate Corpus

Thank you for your Attention



Q/A Part


Questions
Question 1
Question 2
Question 3
Answer 1
Answer 2
Answer 3
Answers