東京工業大学
Tokyo Institute of Technology

OKAZAKILAB

Building a Document-Level Relation Extraction Dataset Assisted by Cross-Lingual Transfer

Youmi Ma, An Wang, Naoaki Okazaki
Tokyo Institute of Technology
{youmi.ma@nlp., an.wang@nlp., okazaki@}c.titech.ac.jp

Outline

Building a DocRE Dataset w. xLingual Transfer

- **Purpose**
  - explore how to build Document-level Relation Extraction (DocRE) datasets with minimal human efforts
- **Method**
  - automatic annotation using cross-lingual transfer -> 😥
  - human annotation assisted by cross-lingual transfer -> 😀
- **Contributions & Findings**
  - collected the first Japanese DocRE dataset
  - showed that although the automatic annotation is not ready for use on its own, it serves as a good start point for human annotation

Outline

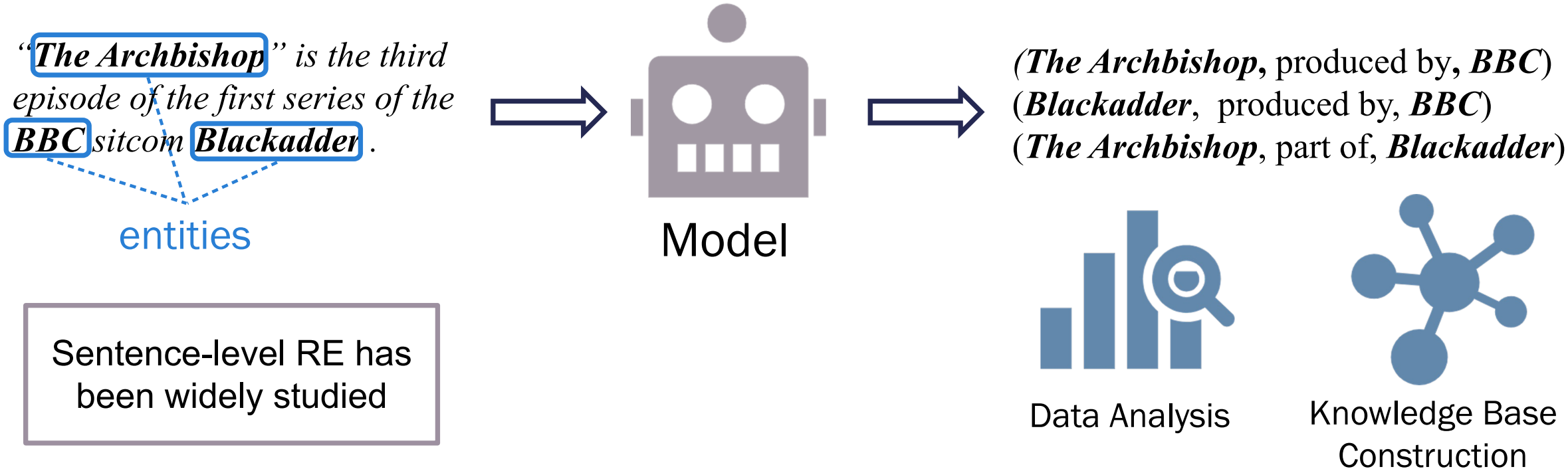
Building a DocRE Dataset w. xLingual Transfer

- **Background & Motivation**
  - task definition: DocRE
  - existing DocRE datasets
- **Approaches**
  - automatic annotation: Re-DocRED^{ja}
  - semi-automatic annotation: JacRED
- **Dataset Analysis & Experiments**
  - superiority of our annotation approach
  - usefulness of our collected dataset

Background

Relation Extraction (RE)

- Task to extract **relations** from natural language texts
  - Identify relations between entity pairs within a sentence



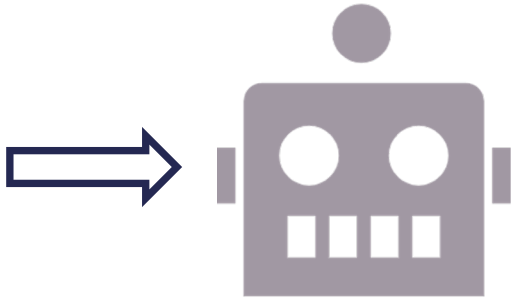
Sentence-level RE has been widely studied

Background

Document-level Relation Extraction (DocRE)

- **Sentence-level RE** is over-simplified (Yao+, 2019)
  - Relations exist beyond sentence boundaries

"***The Archbishop***" is the third episode of the first series of the ***BBC*** sitcom ***Blackadder*** (***The Black Adder***). It is set in ***England*** in ***the late 15th century***, and follows the exploits of the fictitious ***Prince Edmund*** as he is invested as ***Archbishop of Canterbury*** amid a ***Machiavellian*** plot by the King …



Model

(***The Archbishop***, produced by, ***BBC***)
(***Blackadder***, produced by, ***BBC***)
(***The Archbishop***, part of, ***Blackadder***)
(***Prince Edmund***, present in work, ***Blackadder***)
(***Machiavellian***, present in work, ***Blackadder***)
…

cross-sentence relations

Background

Document-level Relation Extraction (DocRE)

- Task to decide relations between **all entity pairs** in a document

challenge

decide relations based on information from the whole document

The Archbishop

[1] "The Archbishop" is the third episode of the first series of the BBC sitcom ***Blackadder*** (***The Black Adder***). [2] It is set in England in the late 15th century, and follows the exploits of the fictitious ***Prince Edmund*** as he is invested as Archbishop of Canterbury amid a Machiavellian plot by the King to acquire lands from the Catholic Church. [3] ... [5] ***Edmund***, faced with the threat of assassination, attempts to escape to France into self-imposed exile; and in a later scene, two drunk knights overhear King Richard IV exclaiming "Who will rid me of this turbulent priest?" [6] The words attributed to King Henry II which led to Becket's death in 1170, and embark on a mission to murder ***Edmund***. [7] …

Subject: ***Prince Edmund*** **Relation**: ***present in work***
Object: ***Blackadder***

Example from DocRED (Yao+, 2019)

Background

Evidence in DocRE(Yao+, 2019)

- **Evidence:** Minimal set of sentences enough for relation decision

challenge

decide relations based on information from the whole document

information filtering

Focus more on **evidence sentences** relevant to current entity pair

The Archbishop

[1] "The Archbishop" is the third episode of the first series of the BBC sitcom ***Blackadder*** (***The Black Adder***). [2] It is set in England in the late 15th century, and follows the exploits of the fictitious ***Prince Edmund*** as he is invested as Archbishop of Canterbury amid a Machiavellian plot by the King to acquire lands from the Catholic Church. [3] ... [5] ***Edmund***, faced with the threat of assassination, attempts to escape to France into self-imposed exile; and in a later scene, two drunk knights overhear King Richard IV exclaiming "Who will rid me of this turbulent priest?" [6] The words attributed to King Henry II which led to Becket's death in 1170, and embark on a mission to murder ***Edmund***. [7] …

Subject: ***Prince Edmund*** **Relation**: ***present in work***
Object: ***Blackadder*** **Evidence: 1,2**

Motivation

Annotating DocRE from Scratch is Difficult

- Heavy burden for annotators affect the quality of collected dataset

"The Archbishop" is the third episode of the first series of the BBC sitcom Blackadder .

Identify all relations in this sentence.

The Archbishop

[1] "The Archbishop" is the third episode of the first series of the BBC sitcom ***Blackadder*** (***The Black Adder***). [2] It is set in England in the late 15th century, and follows the exploits of the fictitious ***Prince Edmund*** as he is invested as Archbishop of Canterbury amid a Machiavellian plot by the King to acquire lands from the Catholic Church. [3] ... [5] ***Edmund***, faced with the threat of assassination, attempts to escape to France into self-imposed exile; and in a later scene, two drunk knights overhear King Richard IV exclaiming "Who will rid me of this turbulent priest?" [6] The words attributed to King Henry II which led to Becket's death in 1170, and embark on a mission to murder ***Edmund***. [7] …

Identify all relations in this document.

Motivation

Existing DocRE Datasets

- Datasets in English, Chinese and Korean are available
  - collected individually despite of high human annotation costs
  - Hinders DocRE research from scaling up

Dataset	Language	# Instances	# Docs	Evidence
DocRED (Yao+, 2019)	en.	50,503	4,051	Y
Re-DocRED (Tan+, 2022)	en.	120,664	4,053	N
HacRED (Cheng+, 2021)	zh.	56,798	7,731	N
HistRED (Yang+, 2023)	kr.	9,965	5,816	Y

The docs are the same, while Re-DocRED added some relation instances

Motivation

Cross-Lingual Transfer (Projection)

- Dataset for sentence-level RE has been successfully created with translation-based cross-lingual transfer
  - Human evaluation ensured the quality of obtained dataset

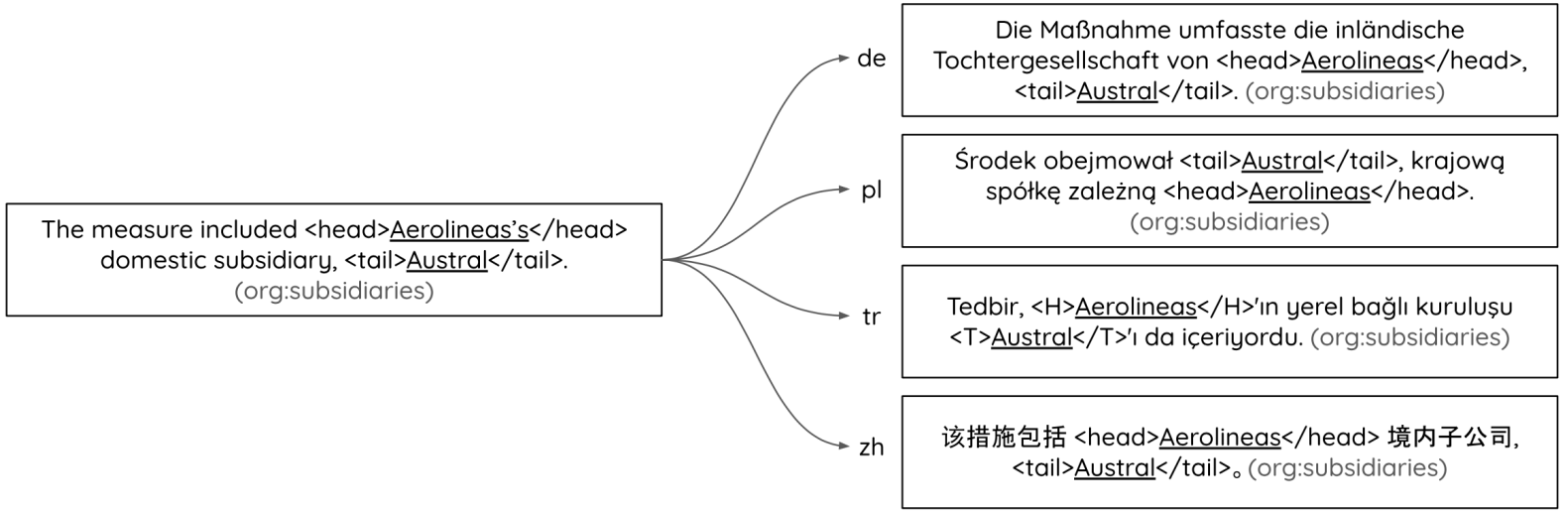


Figure 1: Example translations from English to German, Polish, Turkish and Chinese with XML markup for the head and tail entities to project relation argument annotations.

(Hennig+, 2023)

Motivation

Purpose of This Work

- To explore how DocRE dataset in one language could help collecting DocRE dataset in another using cross-lingual transfer

- To publish a **Japanese** DocRE dataset ready for use

- one of the most widely-used languages for Web content
- one of the most linguistically distant languages from English

Outline

Building a DocRE Dataset w. xLingual Transfer

- **Background & Motivation**
  - task definition: DocRE
  - existing DocRE datasets
- **Approaches**
  - automatic annotation: Re-DocRED^{ja}
  - semi-automatic annotation: JacRED
- **Dataset Analysis & Experiments**
  - superiority of our annotation approach
  - usefulness of our collected dataset

Approach: Automatic Annotation

Starting from English language resources

- **Re-DocRED^{ja}**:Translate Re-DocRED$_{(Tan+, 2022)}$ into Japanese

<e0:LOC> Morogoro Region </e0:LOC> is one of <e1:LOC> Tanzania </e1:LOC> 's <e2:NUM> 31 </e2:NUM> administrative regions .
The regional capital is the municipality of <e3:LOC> Morogoro </e3:LOC> .
According to the <e4:MISC> 2012 national census </e4:MISC> , the region had a population of <e5:NUM> 2,218,492 </e5:NUM> , which was higher than the pre - census projection of <e6:NUM> 2,209,072 </e6:NUM>.

<e0:LOC> モロゴロ州は </e0:LOC>、<e1:LOC> タンザニアに </e1:LOC> ある <e2:NUM> 31 </e2:NUM> 行政区のひとつ。
州都は <e3:LOC> モロゴロ </e3:LOC> 市である。

<e4:MISC> 2012年国勢調査に </e4:MISC> よると、同州の人口は<e5:NUM>2,218,492 </e5:NUM> 人で、国勢調査前の予測値 <e6:NUM> 2,209,072 </e6:NUM> 人を上回った。

Approach: Automatic Annotation

Train a DocRE model on Re-DocRED^{ja}

- The model fails to extract many relations on Japanese Wikipedia

Topic Shift of Contents

Little / No contents about Japanese history / figures / architecture

JA: 堀 直宥((ほり なおさだ、寛文5年11月17日(1665年12月23日) - 正徳元年6月8日(1711年7月23日))は、江戸時代前期から中期の大名で、上総八幡藩第2代藩主。

EN: Naosada Hori (December 23, 1665 - July 23, 1711) was a feudal lord of the early to mid-Edo period, the second lord of the Joso Hachiman domain.

missed relation: (Naosada Hori, head of government, Joso Hachiman domain)

Gap of Surface Structures

Logic of how sentences are organized differs from real Japanese

JA: ザカリアーシュ・ヨージェフ(1924年3月25日 - 1971年11月22日)は、ハンガリー出身のサッカー選手、サッカー指導者。1954年のFIFAワールドカップでは決勝戦を除く4試合にフル出場し準優勝に貢献した。

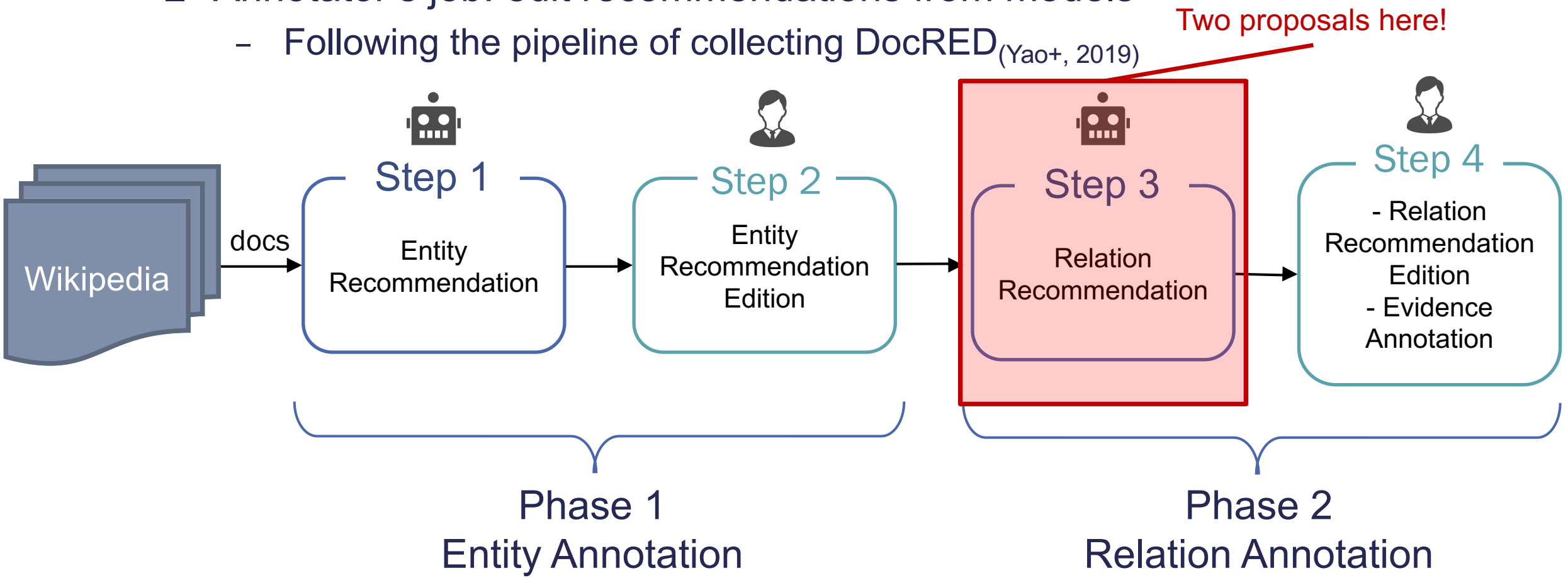
EN: Zakarias Yogev (March 25, 1924 - November 22, 1971) was a Hungarian soccer player and soccer coach. (He) played in all but the final four games of the 1954 FIFA World Cup, contributing to the runners-up finish.

missed relation: (Zakarias Yogev, participant in, the 1954 FIFA World Cup)

Approach: Semi-Automatic Annotation

Pipeline of Human Annotation

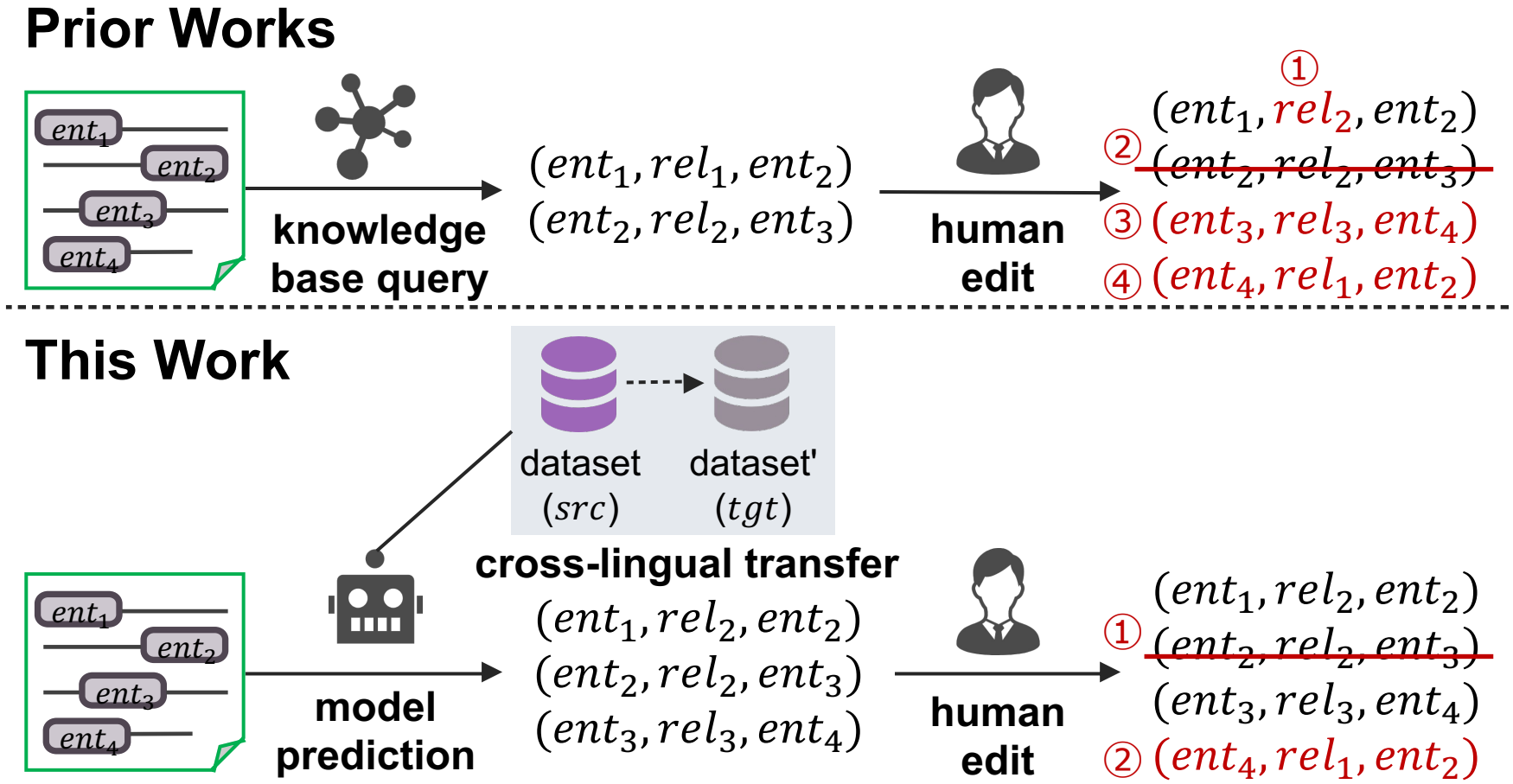
- Annotator's job: edit recommendations from models
  - Following the pipeline of collecting DocRED (Yao+, 2019)



Semi-Automatic Annotation: Proposal 1

Recommend Relations with Model Predictions

- Utilize model trained on translated dataset to recommend relations

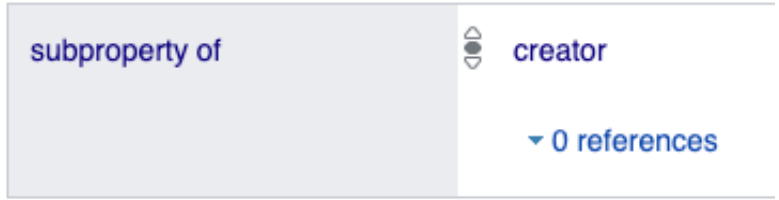


Semi-Automatic Annotation: Proposal 2

Refine Relation Label Set

- Merge labels based on:
  - **Frequency**: select most-frequent relation labels
  - **Hierarchy**:
    - Merge sub-properties into super-properties
      - *E.g. author -> creator*
    - Merge inverse properties
      - *E.g. has part(s) -> part of*
  - **Similarity**: pretrained graph embedding: GraphVite (Zhu+, 2019)

author



has part(s)

- Reduced relation labels from 96 to 35
  - While keeping >88% relation instances in Re-DocRED

Semi-Automatic Annotation

Collected Dataset: JacRED

- **Ja**panese Do**c**ument-level **R**elation **E**xtraction **D**ataset

Dataset	Language	# Instances	# Docs	Evidence
DocRED (Yao+, 2019)	en.	50,503	4,051	Y
Re-DocRED (Tan+, 2022)	en.	120,664	4,053	N
HacRED (Cheng;, 2021)	zh.	56,798	7,731	N
HistRED (Yang+, 2023)	kr.	9,965	5,816	Y
JacRED	ja.	42,241	2,000	Y

Outline

Building a DocRE Dataset w. xLingual Transfer

- **Background & Motivation**
  - task definition: DocRE
  - existing DocRE datasets
- **Approaches**
  - automatic annotation: Re-DocRED^{ja}
  - semi-automatic annotation: JacRED
- **Dataset Analysis & Experiments**
  - superiority of our annotation approach
  - usefulness of our collected dataset

Dataset Analysis

Statistics: Comparison with Existing Datasets

- JacRED has its advantages over existing datasets as a general language resource
  - Even without the distinctiveness of language

 # Sentences, **# Entities** and **# Relations** averaged over documents
 # Evidences averaged over relation instances

	# Sentences	# Entities	# Relations	# Evidences
DocRED$_{(Yao+, 2019)}$	7.98	19.51	12.45	1.60
Re-DocRED$_{(Tan+, 2022)}$	7.98	19.45	29.77	0.88
JacRED	8.39	17.87	17.87	1.67

more relation instances

more evidence instances

Dataset Analysis

Statistics: Number of Human Edits

- Sample 400 documents from JacRED and calculate the number of edit steps before reaching at final annotations
  - Starting from model predictions reduces the number of human edit steps

	# Recommends.	# Deletions	# Substitutions	# Additions
Model Predictions	**6,500**	**1,266**	**224**	**2,740**
Knowledge Base Queries	3,200	1,459	113	6,233

more recommendations

Fewer instances to add

Experiments

Training on Translated Dataset

- Evaluate DREEAM$_{(Ma+, 2023)}$ trained with Re-DocRED^{ja} on JacRED
  - Number of documents used during training indicated in parenthesis

	Precision	Recall	F1
JacRED (1,400)	**64.76**	**73.29**	**68.73**
Re-DocRED^{ja} (3,053)	56.14	53.67	54.87

Models trained on Re-DocRED^{ja} suffers from low recall.

Experiments

Training on Translated Dataset

- Evaluate DREEAM$_{(Ma+, 2023)}$ trained with Re-DocRED^{ja} on JacRED
  - Number of documents used during training indicated in parenthesis

	Precision	Recall	F1
JacRED (1,400)	**64.76**	**73.29**	**68.73**
Re-DocRED^{ja} (3,053)	56.14	53.67	54.87
Re-DocRED^{ja} (1,400)	55.52	51.77	53.56

Models trained on Re-DocRED^{ja} suffers from low recall.

The performance gap between models trained on Re-DocRED^{ja} and JacRED is evident under the same setting.

Summary

Building a DocRE Dataset w. xLingual Transfer

dataset

- **Purpose**
  - explore how to build Document-level Relation Extraction (DocRE) datasets with minimal human efforts
- **Method**
  - automatic annotation using cross-lingual transfer -> 😥
  - human annotation assisted by cross-lingual transfer -> 😀
- **Contributions & Findings**
  - collected the first Japanese DocRE dataset
  - showed that although the automatic annotation is not ready for use on its own, it serves as a good start point for human annotation

References

- Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. **DocRED: A Large-Scale Document-Level Relation Extraction Dataset.** ACL 2019.
- Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. **HacRED: A Large-Scale Relation Extraction Dataset Toward Hard Cases in Practical Applications**. ACL-IJCNLP 2021 Findings.
- Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. **HistRED: A Historical Document-Level Relation Extraction Dataset**. ACL 2023.
- Leonhard Hennig, Philippe Thomas, and Sebastian Möller. **MultiTACRED: A Multilingual Version of the TAC Relation Extraction Dataset**. ACL 2023.
- Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, Sharifah Mahani Aljunied. **Revisiting DocRED -- Addressing the False Negative Problem in Relation Extraction**. EMNLP 2022.
- Youmi Ma, An Wang, Naoaki Okazaki. **DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction**. EACL 2023.
- Zhaocheng Zhu, Shizhen Xu, Meng Qu, Jian Tang. **GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding**. WWWC 2019.
- Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. **Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation**. ACL 2022 Findings.
- Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. **Document-level relation extraction with adaptive thresholding and localized context pooling**. AAAI 2021.
- Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. **Document- level relation extraction as semantic segmentation**. IJCAI 2021.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 2022.