

ezCoref

Towards Unifying Annotation Guidelines for Coreference Resolution

Ankita Gupta Marzena Karpinska Wenlong Zhao Kalpesh Krishna
Jack Merullo Luke Yeh Mohit Iyyer Brendan O'Connor

UMassAmherst

Manning College of Information
& Computer Sciences

Coreference Resolution

Background

Determine which spans of text refer to the same entity

[John] doesn't like [Fred], but [he] still invited [him] to [the party].

Coreference Resolution

Background

Determine which spans of text refer to the same entity

[John] doesn't like [Fred], but [he] still invited [him] to [the party].

Antecedent ← Mention

Coreference Resolution

Background

Several datasets have been collected to facilitate modeling of coreferences.



OntoNotes, ARRAU



LitBank, GUM, Phrase Detectives

and many more...

Coreference Resolution

Motivation

However, application to other domains and languages requires new dataset curation.



Legal



Healthcare



Finance



Coreference Resolution

Motivation

To collect new datasets we need

1. Annotation guidelines



Coreference Resolution

Motivation

To collect new datasets we need

1. Annotation guidelines



2. Annotation Tool



Coreference Resolution

Motivation

To collect new datasets we need

1. Annotation guidelines



2. Annotation Tool



3. Annotation Workforce



Coreference Resolution

Motivation

To collect new datasets we need

1. Annotation guidelines



2. Annotation Tool



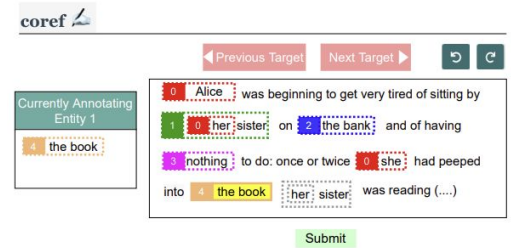
3. Annotation Workforce



Existing datasets differ widely in these aspects
making **new annotation efforts challenging**

Key Contributions

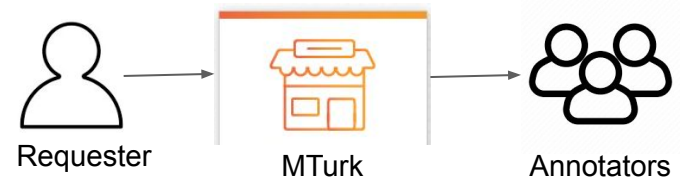
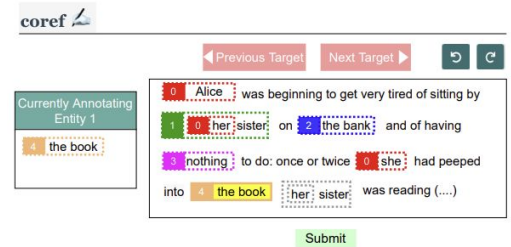
1. **ezCoref annotation tool**
 - a. Open-sourced
 - b. Integration with crowdsourcing platforms
 - c. Interactive Tutorial



Key Contributions

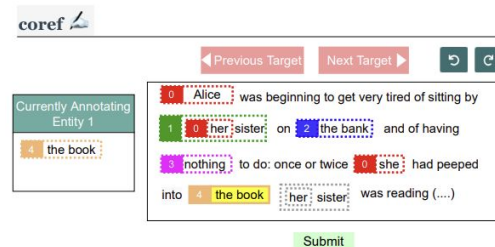
1. ezCoref annotation tool
 - a. Open-sourced
 - b. Integration with crowdsourcing platforms
 - c. Interactive Tutorial

2. **Re-annotation study**
 - a. seven coreference datasets
 - b. seven domains



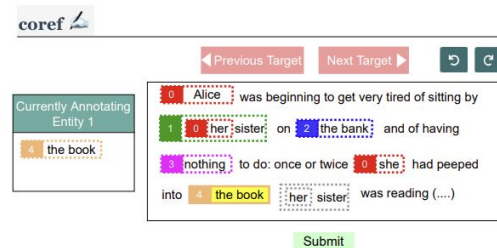
Key Contributions

1. ezCoref annotation tool
 - a. Open-sourced
 - b. Integration with crowdsourcing platforms
 - c. Interactive Tutorial
2. Re-annotation study
 - a. seven coreference datasets
 - b. seven domains
3. **Annotation analysis**
 - a. ezCoref vs. ezCoref annotations
 - b. ezCoref vs. original annotations



Key Contributions

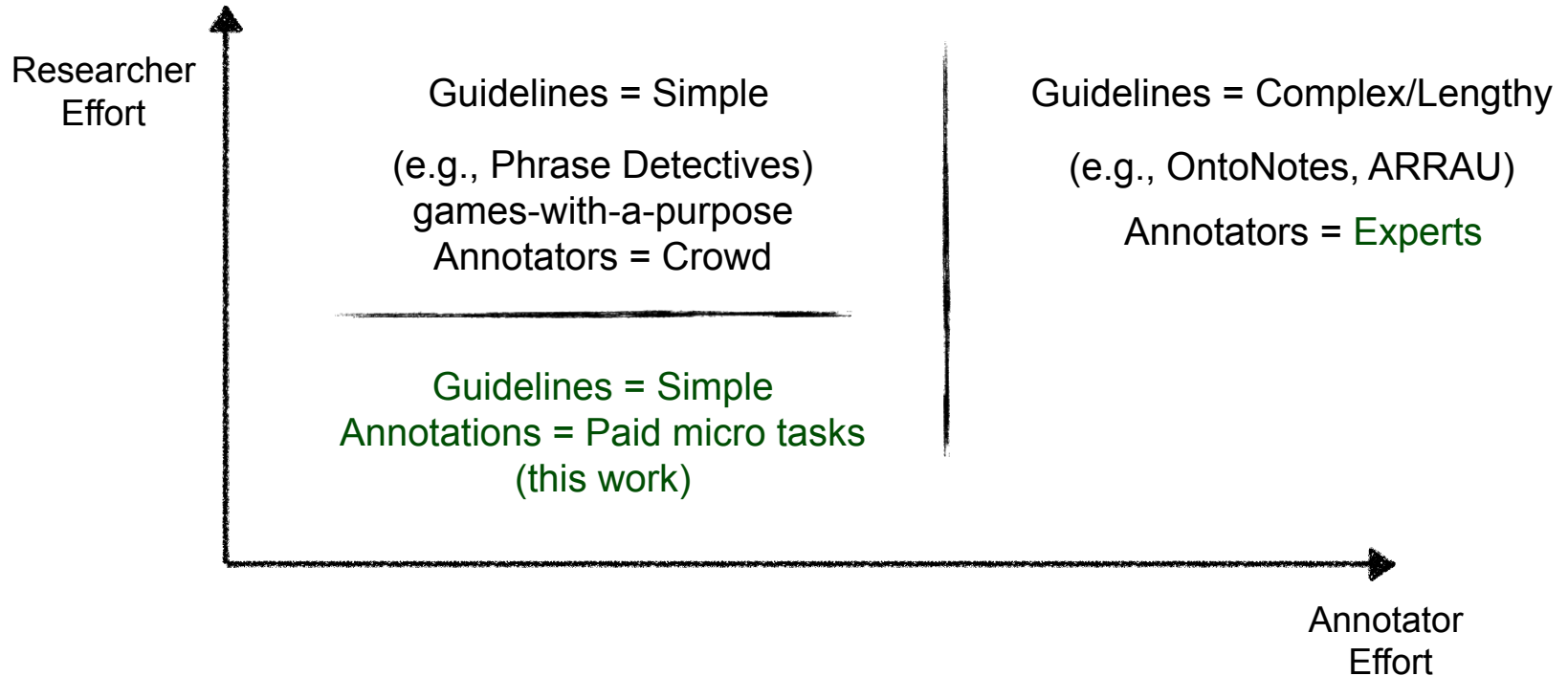
1. ezCoref annotation tool
 - a. Open-sourced
 - b. Integration with crowdsourcing platforms
 - c. Interactive Tutorial
2. Re-annotation study
 - a. seven coreference datasets
 - b. seven domains
3. **Annotation analysis**
 - a. agreements/disagreements among crowd annotators
 - b. comparison with original annotations



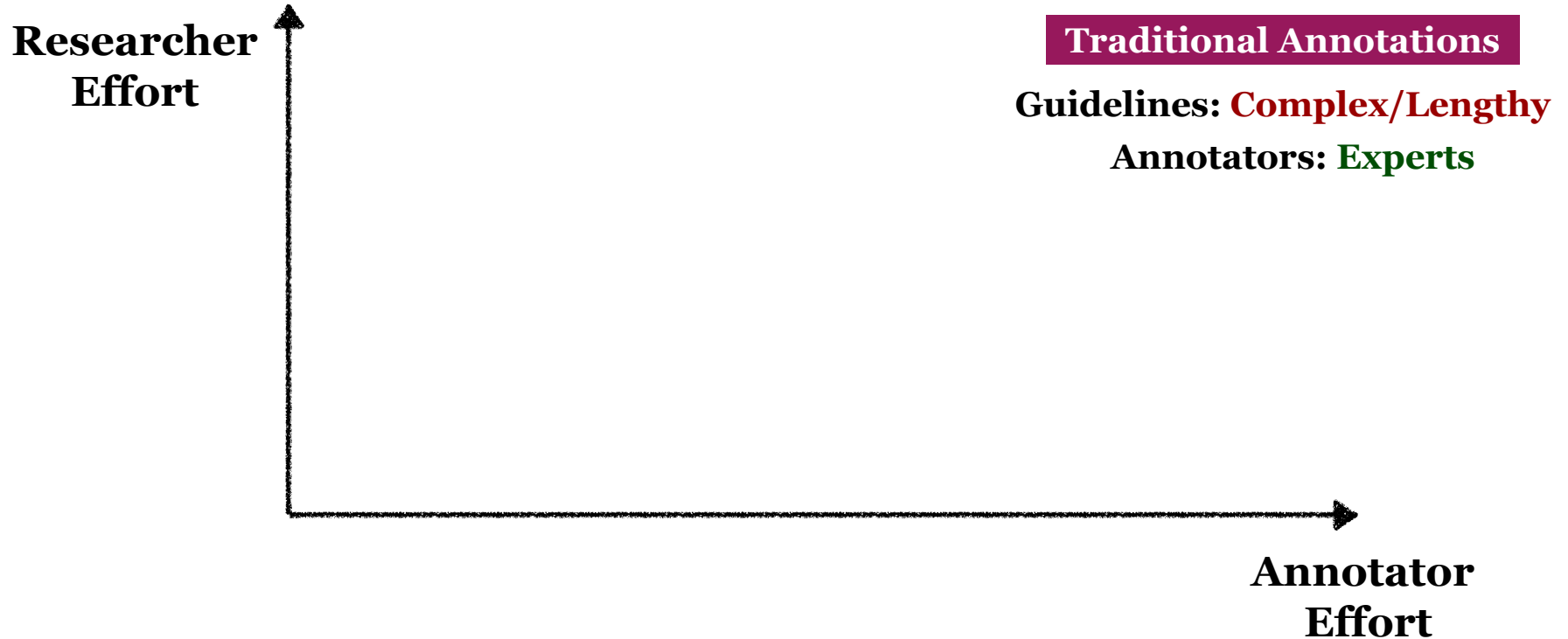
1. ezCoref Annotation Tool

Annotator population and annotation platforms

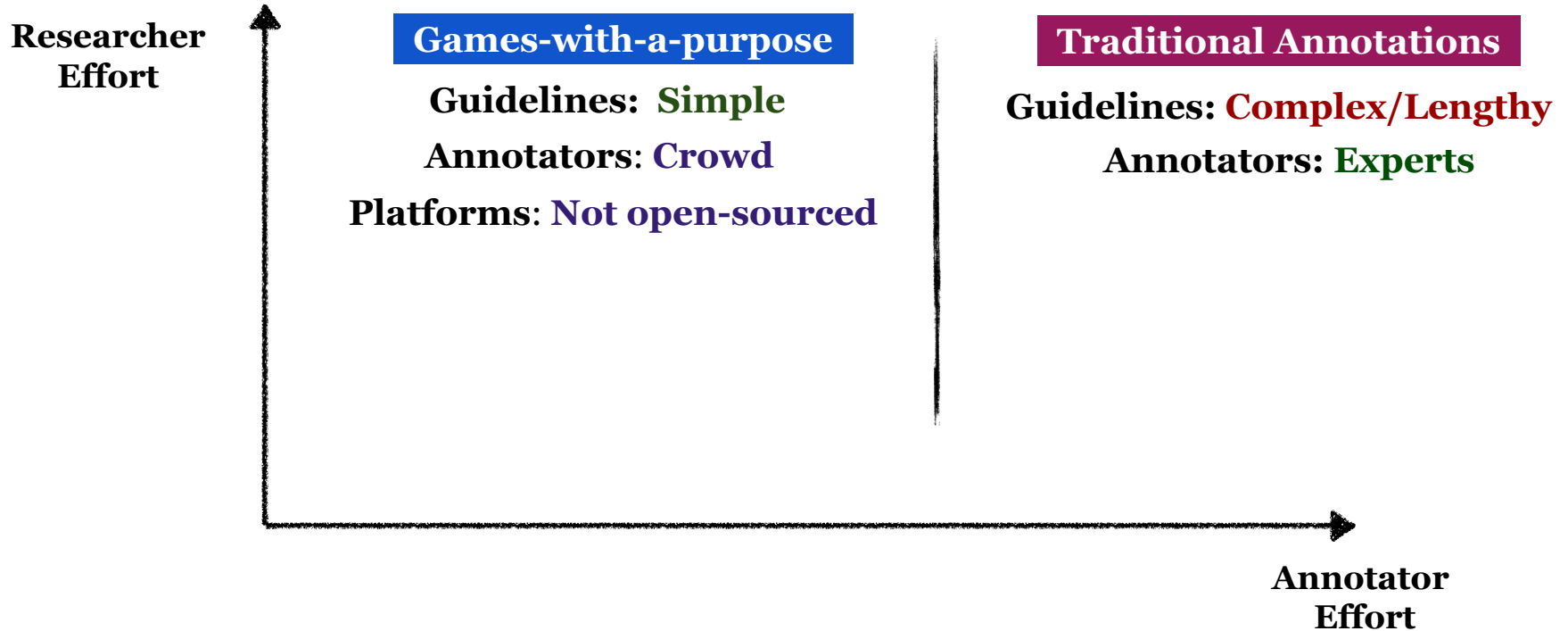
Background



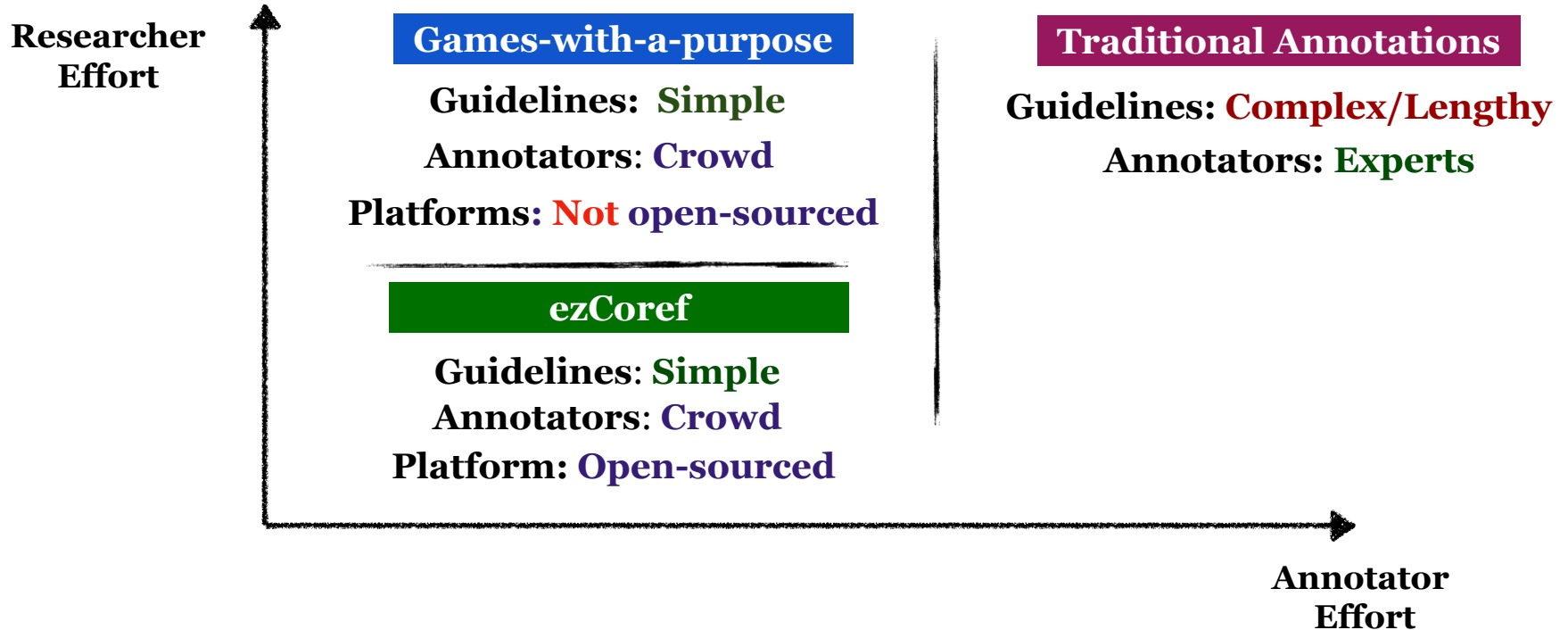
Annotation Approaches



Annotation Approaches



Annotation Approaches



ezCoref

Annotation Tool and Tutorial



Open Sourced



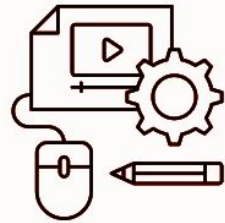
**Integration with
crowdsourcing platforms
(e.g., AMT, crowdflower)**



**Crowdsourcing
Friendly UI**

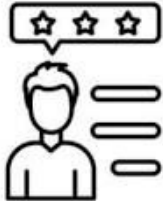
ezCoref

Annotation Tool and Tutorial



Tutorial

- ✓ Interactive
- ✓ Customizable
- ✓ Validate by 70+ AMT annotators



AMT annotator

“Absolutely beautiful, intuitive, and helpful. Legitimately the best one I’ve ever seen in my 2 years on AMT! Awesome job”

ezCoref

Annotation Tool and Tutorial



2. Reannotation Study

Crowdsourcing Setup

OntoNotes

Hovy et al.
2006

GUM

Zeldes
2017

ARRAU

Uryupina et al.
2019

LitBank

Bamman et al.
2020

QuizBowl

Guha et al.
2015

Phrase Detectives

Chamberlain et al.
2016

PreCo

Chen et al.
2018

Linguistic Experts

**Domain
Experts**

Crowd

Crowdsourcing Setup

OntoNotes

Hovy et al.
2006

GUM

Zeldes
2017

ARRAU

Uryupina et al.
2019

LitBank

Bamman et al.
2020

QuizBowl

Guha et al.
2015

Phrase Detectives

Chamberlain et al.
2016

PreCo

Chen et al.
2018

Linguistic Experts

Domain Experts

Crowd



News



Biographies



Weblogs



Fiction



Quizzes



Opinions



Wikipedia

Crowdsourcing Setup

OntoNotes

Hovy et al.
2006

GUM

Zeldes
2017

ARRAU

Uryupina et al.
2019

LitBank

Bamman et al.
2020

QuizBowl

Guha et al.
2015

Phrase Detectives

Chamberlain et al.
2016

PreCo

Chen et al.
2018

Linguistic Experts

Domain
Experts

Crowd



News



Biographies



Weblogs



Fiction



Quizzes



Opinions



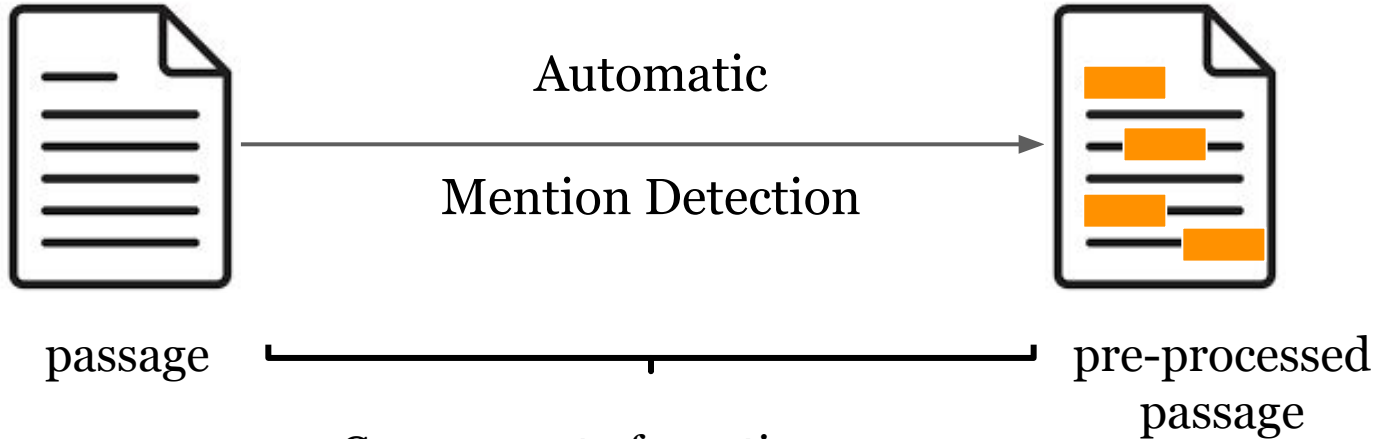
Wikipedia



240 passages

Crowdsourcing Setup

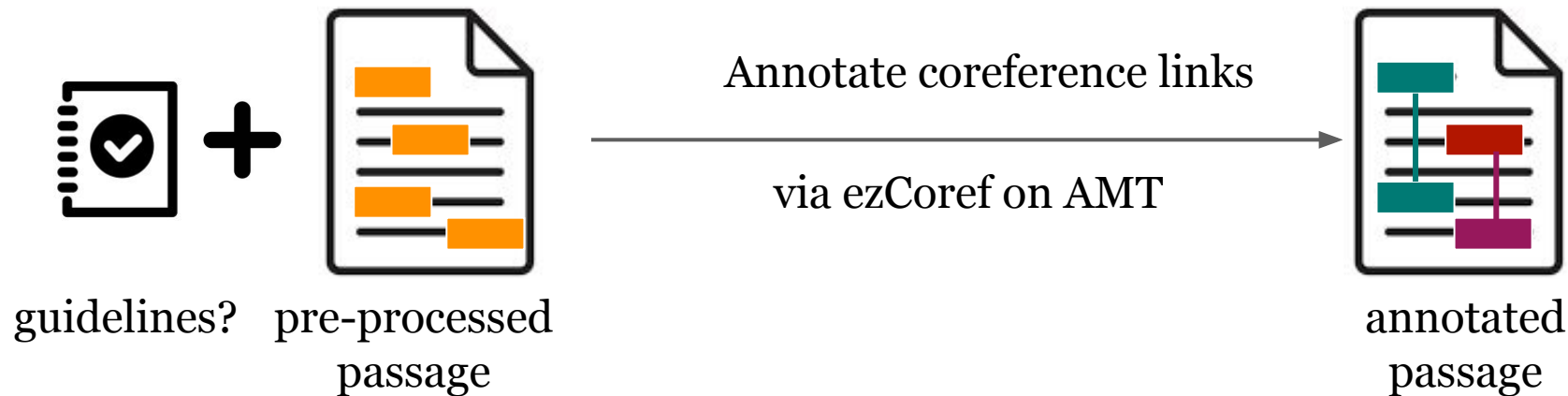
Preprocessing



- Common set of mentions
- High recall (>80%) across datasets

Crowdsourcing Setup

Annotating Coreferences



Annotation Guidelines

Inconsistencies

Maybe we need a CIA version of the Miranda warning: You have the right to conceal your coup intentions, because we may rat on you.

An example sentence drawn from Wall Street Journal

Annotation Guidelines

Inconsistencies

Maybe we need a CIA version of the Miranda warning: You have the right to conceal your coup intentions, because we may rat on you.

OntoNotes

Does not mark generic pronouns

Annotation Guidelines

Inconsistencies

Maybe we need a CIA version of the Miranda warning: You have the right to conceal your coup intentions, because we may rat on you.

OntoNotes

Does not mark generic pronouns

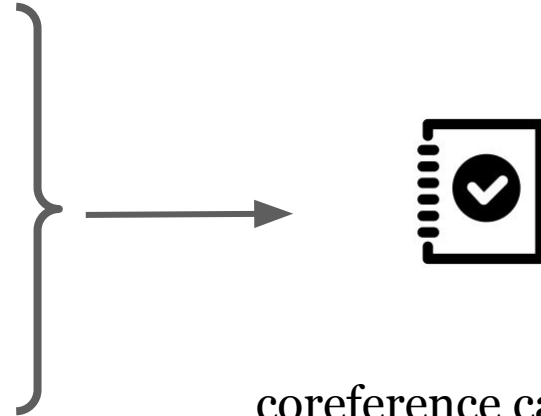
Maybe [we] need a CIA version of the Miranda warning: [You] have the right to conceal [your] coup intentions, because [we] may rat on [you].

ARRAU

Marks generic pronouns as *undef-reference* (not coreferent)

Annotation Guidelines

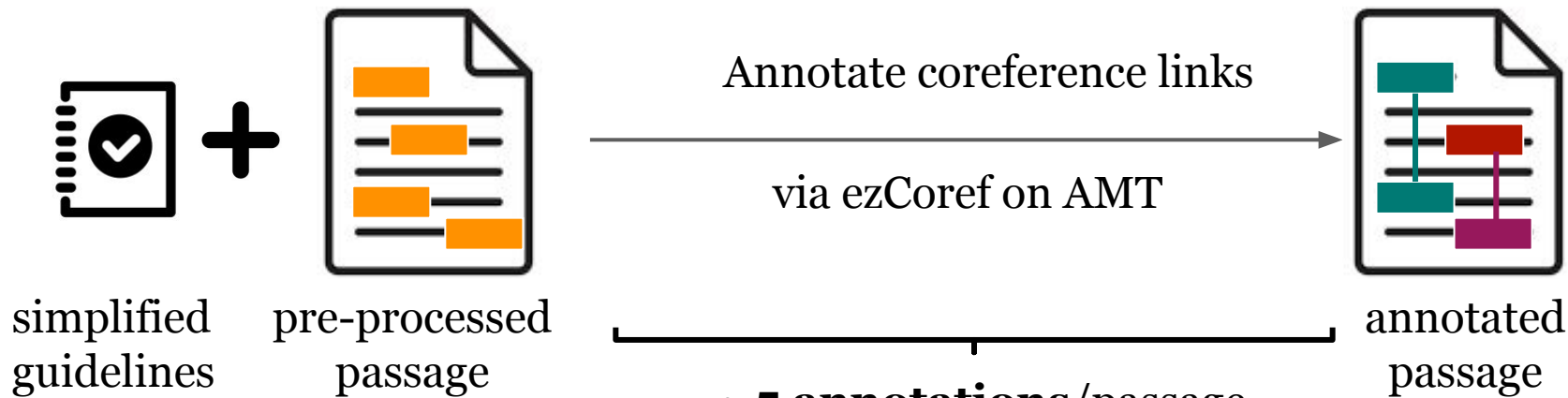
RQ: What **guidelines** to use when collecting new annotations?



coreference cases treated
uniformly across all datasets
as **simplified guidelines**

Crowdsourcing Setup

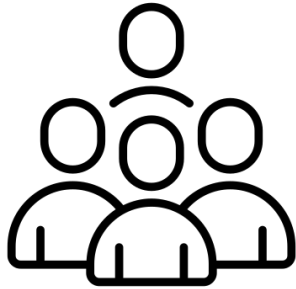
Annotating Coreferences



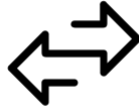
- **5 annotations**/passage
- **Majority voting**
- **12,200** mentions
- **42,108** tokens

3. Analysis of ezCoref annotations

Investigating Agreements



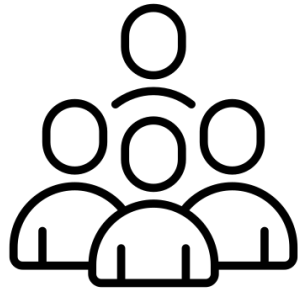
ezCoref Annotations



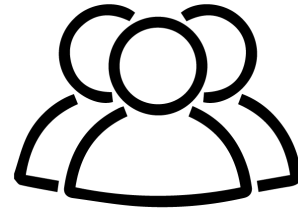
Original Annotations

1. **Comparison** of ezCoref annotations with original annotations

Investigating Agreements



ezCoref Annotations

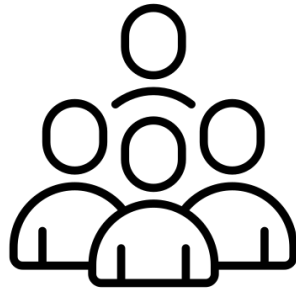


Original Annotations

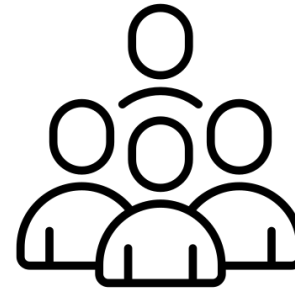
(linguists, domain experts, crowd)

1. **Comparison** of ezCoref annotations with original annotations

Investigating Agreements



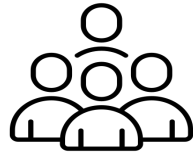
ezCoref Annotations



ezCoref Annotations

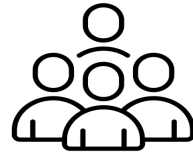
2. Agreements and disagreements among ezCoref annotators

Investigating Agreements

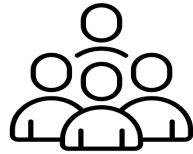


ezCoref Annotations

=



ezCoref Annotations



ezCoref Annotations

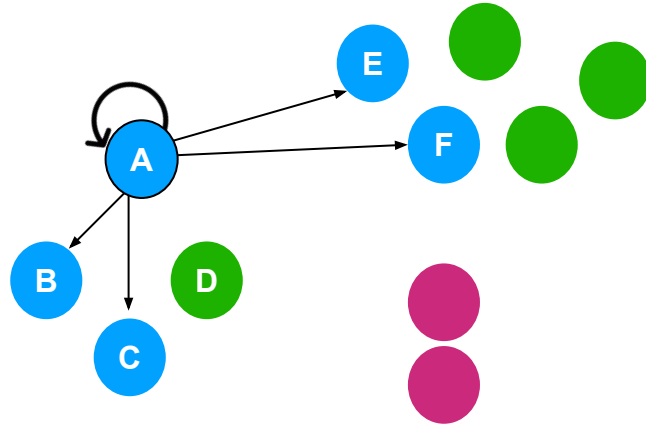
≠



Original Annotations

3. Deviations of ezCoref annotations from original annotations

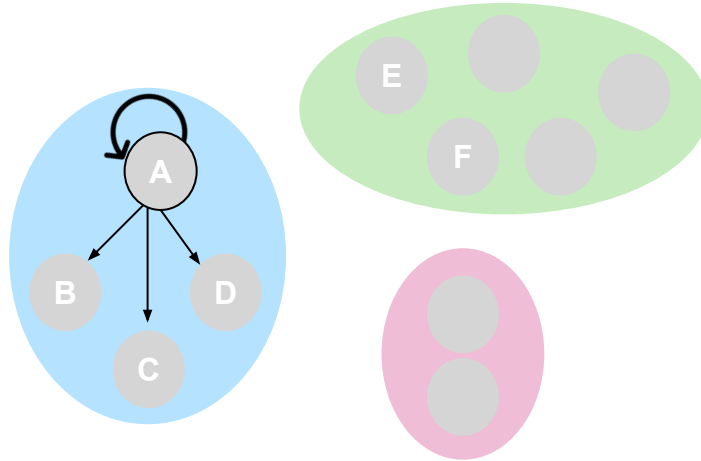
Measuring Agreements



Mention level annotations by annotator 1



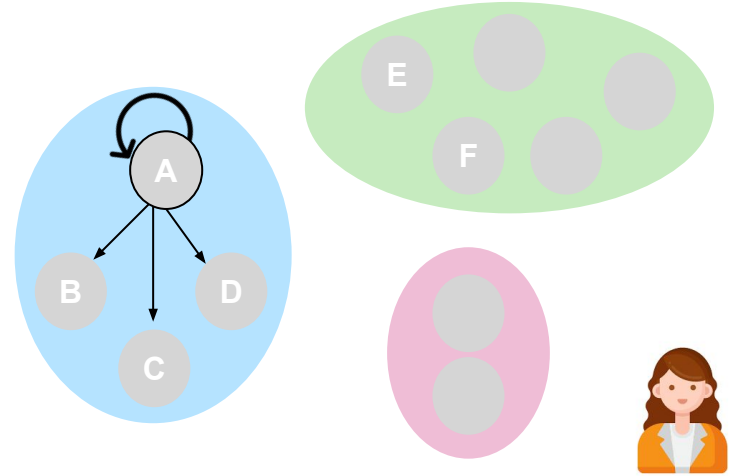
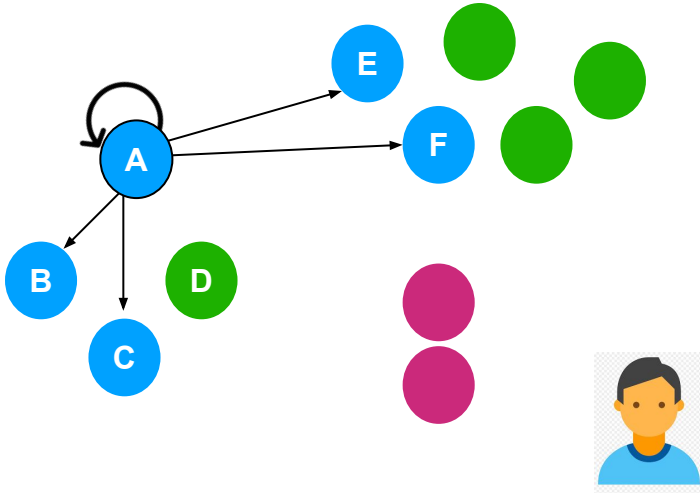
Measuring Agreements



Mention level annotations by annotator 2

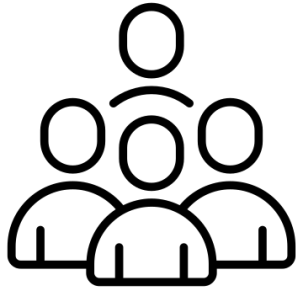


Measuring Agreements



Agreement: Fraction of mentions two annotators agree should be coreferent with a given mention, which is captured by **B3 precision and recall measure**

Investigating Agreements



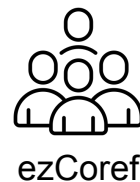
ezCoref Annotations



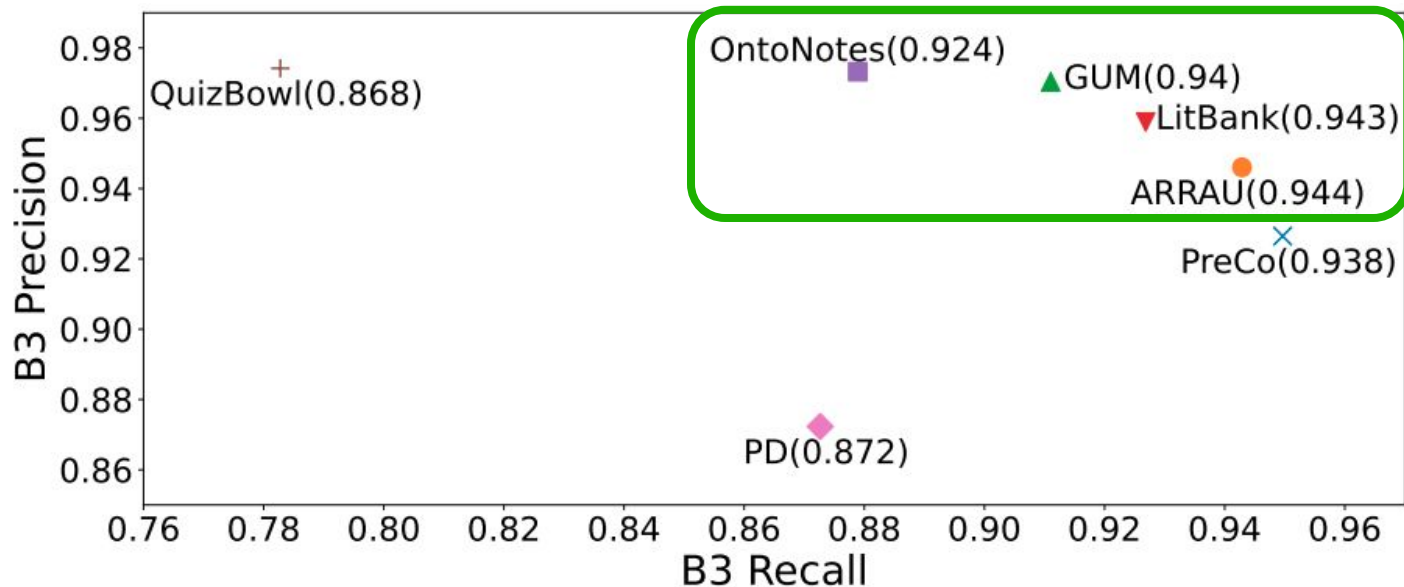
Original Annotations

1. **Comparison** of ezCoref annotations with original annotations

Analysis



1. Comparison of ezCoref and original annotations

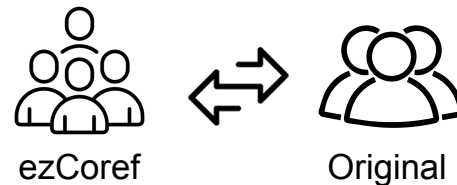


High agreement with expert annotated datasets

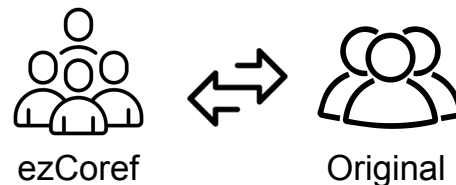
Analysis

High agreement with expert annotated datasets

- **Pronouns**
- **Named Entities**



Analysis



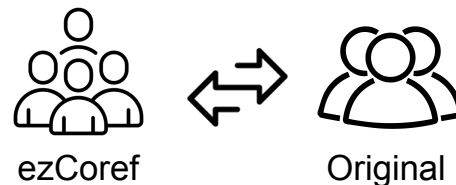
High agreement with expert annotated datasets

- Pronouns
- Named Entities
- Appositives**

[Fuad Basya]_{e1}, [spokesman for the Indonesian military]_{e1}, said fisherman first noticed the people and a warship was deployed to retrieve them.

An example of **appositive construction** marked as **coreferent** by **ezCoref** annotators, consistent with the **GUM** guidelines

Analysis



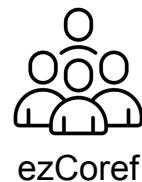
High agreement with expert annotated datasets

- Pronouns
- Named Entities
- Appositives
- Generic mentions**

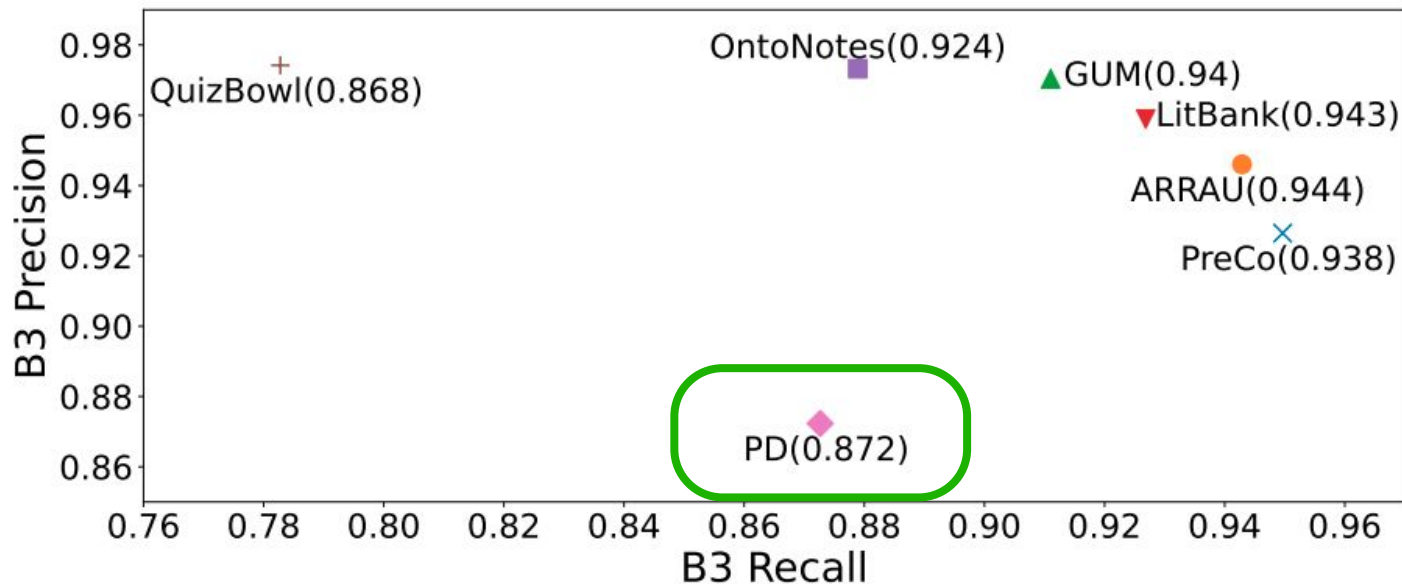
Kidder is putting [brokers] through a 20 week training course (..) However, skeptics caution (..) [brokers] and investment bankers looks great on paper, but doesn't always happen.

An example of **generic mentions** annotated as **coreferent** by **ezCoref** annotators, **consistent** with the **ARRAU** guidelines

Analysis

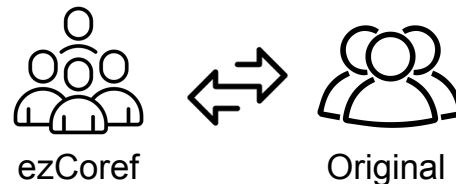


1. Comparison of ezCoref and original annotations



Low precision with Phrase Detectives suggesting ezCoref annotators merge original clusters

Analysis

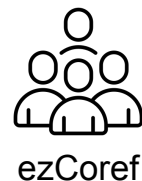


Low precision with Phrase Detectives suggesting ezCoref annotators merge original clusters

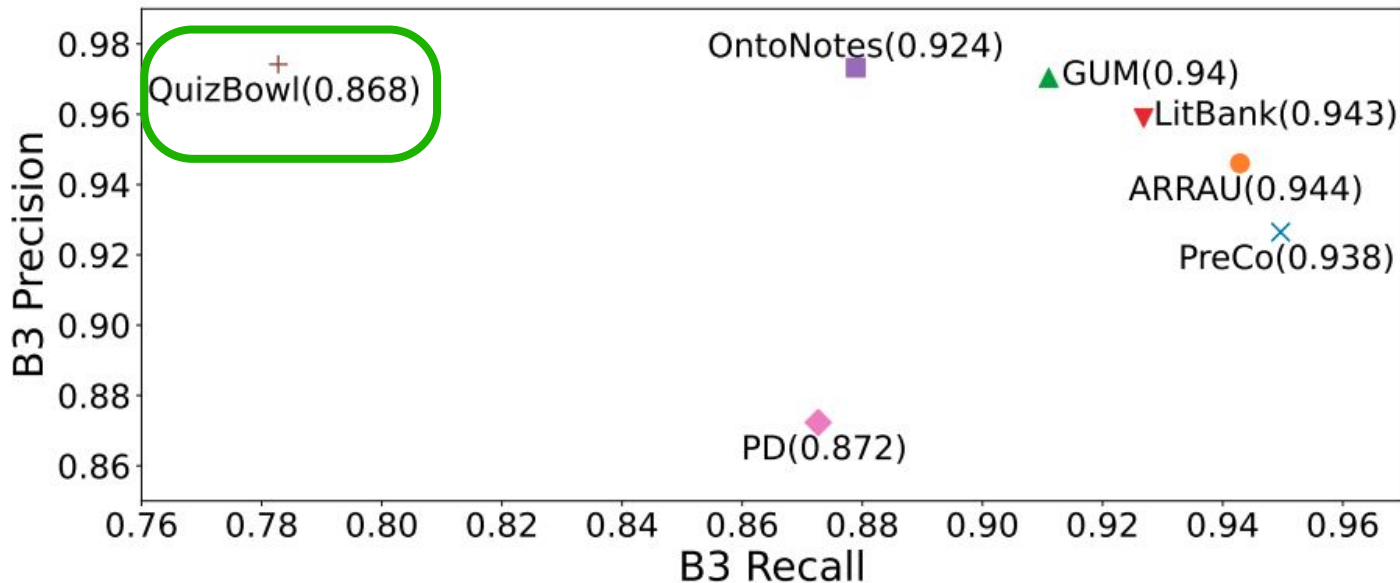
Not long after **[a suitor]** appeared (..) and the miller (..) betrothed his daughter to **[him]**. But the girl did not care for **[the man]**(...) she could not look at **[him]** nor think of **[him]**; without an inward shudder.

An example of **split-entities (missing links)** in Phrase Detectives' annotations. Instead, **ezCoref annotators** mark all mentions as **referring to the same entity**.

Analysis

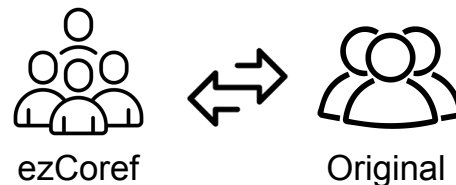


1. Comparison of ezCoref and original annotations



Low recall with QuizBowl suggesting ezCoref annotators miss mentions in the original clusters

Analysis



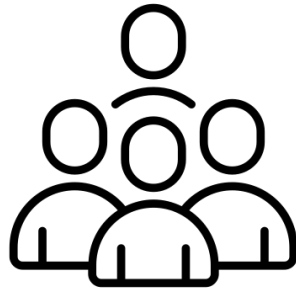
Low recall with QuizBowl suggesting ezCoref annotators miss mentions in the original clusters

[One character in this work]; is forgiven by [his] wife for an affair with a governess before beginning one with a ballerina.

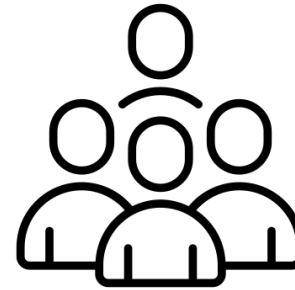
In addition to [Stiva]; and [Nikolai], [another character in this work] (...) had earlier failed in [his] courtship of Ekaterina Shcherbatskaya.

An example from **QuizBowl** dataset with **cataphoric references missed** by **ezCoref** annotators

Investigating Agreements



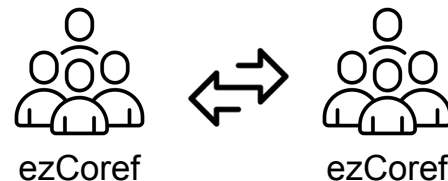
ezCoref Annotations



ezCoref Annotations

2. Agreements and disagreements among ezCoref annotators

Analysis

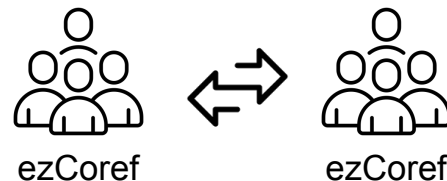


2. Agreements among ezCoref annotators

Which domains are suitable for crowdsourcing?

Fiction	Biographies	Opinion	Web	News	Wiki	Quiz
----------------	--------------------	----------------	------------	-------------	-------------	-------------

Analysis



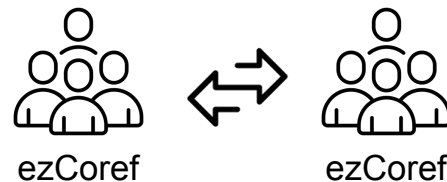
2. Agreements among ezCoref annotators

Which domains are suitable for crowdsourcing?

Fiction	Biographies	Opinion	Web	News	Wiki	Quiz
72.6	72.4	69.5	65.9	62.3	61.8	59.7

Domain-wise Inter Annotator **Agreement** using B3 (F1) scores

Analysis



2. Agreements among ezCoref annotators

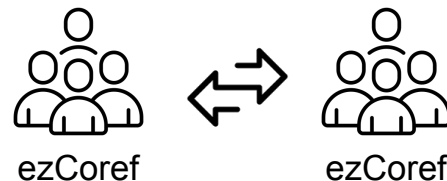
Which domains are suitable for crowdsourcing?

Fiction	Biographies	Opinion	Web	News	Wiki	Quiz
72.6	72.4	69.5	65.9	62.3	61.8	59.7

Domain-wise Inter Annotator **Agreement** using B3 (F1) scores

Fiction domain has highest inter-annotator agreement

Analysis

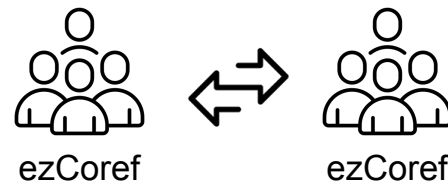


2. Agreements among ezCoref annotators

Fiction domain has highest inter-annotator agreement
abundant in **pronouns**, familiar **childhood stories**

A Wolf had been gorging on an animal **[he]** had killed, when suddenly a small bone in the meat stuck in **[his]** throat and **[he]** could not swallow it. **[He]** soon felt a terrible pain in **[his]** throat (...) **[He]** tried to induce everyone **[he]** met to remove the bone. “**[I]** would give anything,” said **[he]**,” if [you]; would take it out.

Analysis



2. **Agreements** among ezCoref annotators

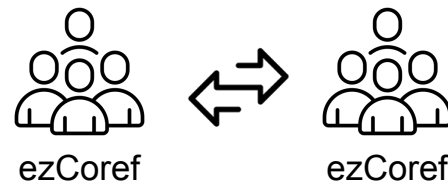
Which domains are suitable for crowdsourcing?

Fiction	Biographies	Opinion	Web	News	Wiki	Quiz
72.6	72.4	69.5	65.9	62.3	61.8	59.7

Domain-wise Inter Annotator **Agreement** using B3 (F1) scores

Quiz domain has lowest inter-annotator agreement

Analysis



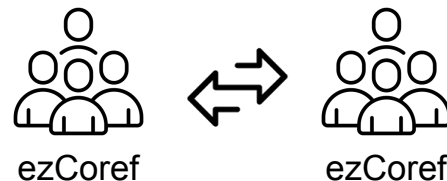
2. Agreements among ezCoref annotators

Quiz domain has lowest inter-annotator agreement

Abundant with **cataphora** and requires **factual knowledge**

[Another character in this work] rejects [Ekaterina] before (...) moving to St. Petersburg. For 10 points name this work in which [Levin] marries [Kitty], (...) a novel by Leo Tolstoy.

Analysis



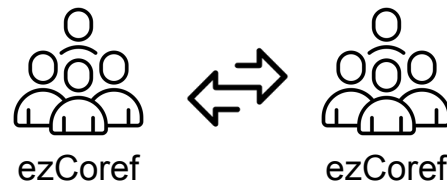
2. Disagreements among ezCoref annotators

Cases of genuine ambiguity in ezCoref annotations

[Fog] everywhere. **[Fog]** up **[the river]**, where **[it]** flows among green aits and meadows; **[fog]** down **[the river]**, where **[it]** rolls defiled among the tiers of shipping and the waterside pollutions of a great (and dirty) city.

—from *Bleak House* by Charles Dickens

Analysis



2. Disagreements among ezCoref annotators

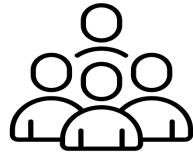
Cases of genuine ambiguity in ezCoref annotations

[Fog] everywhere. **[Fog]** up **[the river]**, where **[it]** flows among green aits and meadows; **[fog]** down **[the river]**, where **[it]** rolls defiled among the tiers of shipping and the waterside pollutions of a great (and dirty) city.

—from *Bleak House* by Charles Dickens

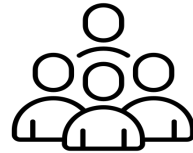
- Disagreement on mention **[it]** among ezCoref annotators
- Consistent with Szakolczai (2016)'s literary analysis

Investigating Agreements

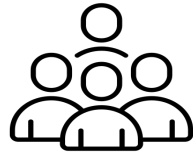


ezCoref Annotations

=



ezCoref Annotations



ezCoref Annotations

≠

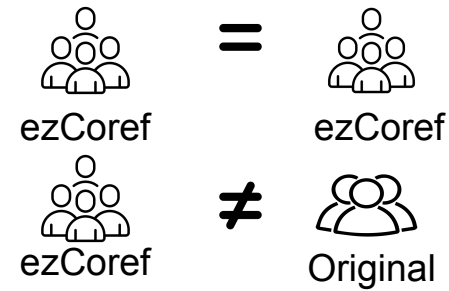


Original Annotations

3. Deviations of ezCoref annotations from original annotations

Analysis

3. Deviations of ezCoref annotations from original annotations



Generic Pronouns

Maybe we need a CIA version of the Miranda warning: You have the right to conceal your coup intentions, because we may rat on you.

OntoNotes

Not marked

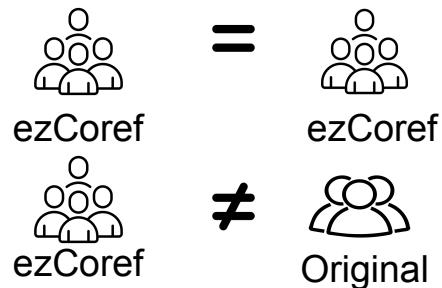
Maybe [we] need a CIA version of the Miranda warning: [You] have the right to conceal [your] coup intentions, because [we] may rat on [you].

ARRAU

Marked as *undef-reference* (not coreferent)

Analysis

3. Deviations of ezCoref annotations from original annotations



Generic Pronouns

Maybe we need a CIA version of the Miranda warning: You have the right to conceal your coup intentions, because we may rat on you.

OntoNotes

Not marked

Maybe [we] need a CIA version of the Miranda warning: [You] have the right to conceal [your] coup intentions, because [we] may rat on [you].

ARRAU

Marked as *undef-reference* (not coreferent)

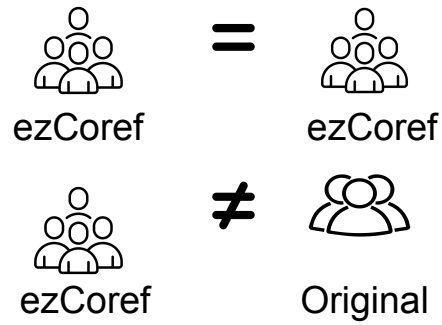
Maybe [we] need a CIA version of the Miranda warning: [You] have the right to conceal [your] coup intentions, because [we] may rat on [you].

ezCoref annotations

Marked as coreferent

Analysis

3. Deviations of ezCoref annotations from original annotations



Appos

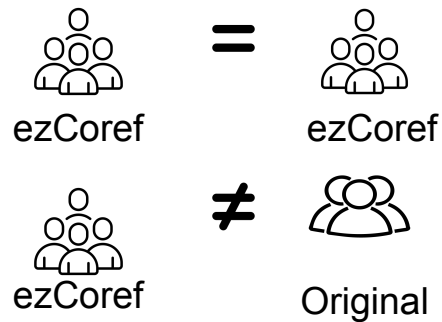
Appositives

[Fuad Basya], [spokesman for the Indonesian military], said (..) a warship was deployed to retrieve them.

LitBank

Analysis

3. Deviations of ezCoref annotations from original annotations



Appos

Appositives

Coref

[Fuad Basya], [spokesman for the Indonesian military], said (..) a warship was deployed to retrieve them.

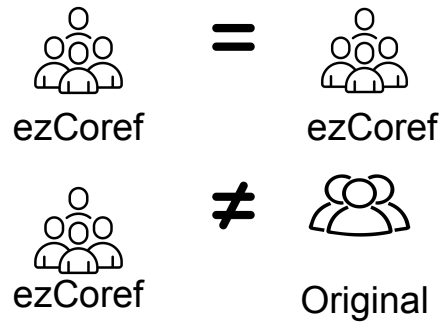
LitBank

[Fuad Basya], [spokesman for the Indonesian military], said (..) a warship was deployed to retrieve them.

GUM

Analysis

3. Deviations of ezCoref annotations from original annotations



Appos

Appositives

Coref

[Fuad Basya], [spokesman for the Indonesian military], said (..) a warship was deployed to retrieve them.

[Fuad Basya], [spokesman for the Indonesian military], said (..) a warship was deployed to retrieve them.

LitBank

GUM

Coref

[Fuad Basya], [spokesman for the Indonesian military], said (..) a warship was deployed to retrieve them.

ezCoref annotations

Takeaways

1. Open-source ezCoref annotation tool, conduct an annotation study
2. Coreference annotation remains a difficult task with **genuine ambiguities** (Plank, 2022)
3. **High-agreement with expert annotated datasets**, suggesting high quality annotations are achievable without extensive training.
4. Interesting **deviations** of ezCoref annotations from original annotations (**e.g., generic pronouns**)
 - a. Research community should revisit these phenomena when curating future unified annotation guidelines

ezCoref

Towards Unifying Annotation Guidelines for
Coreference Resolution



<https://github.com/gnkitaa/ezCoref>

Thank You. Questions?