Towards Evaluation of Cross-document Coreference Resolution Models Using Datasets with Diverse Annotation Schemes

Anastasia Zhukova[1], Felix Hamborg[2], Bela Gipp[1,3]

[1]University of Wuppertal, Germany
[2]University of Konstanz, Germany
[3]University of Göttingen, Germany

DATA & KNOWLEDGE ENGINEERING

LREC'22, 21-23 June 2022, Marseille

BERGISCHE UNIVERSITÄT WUPPERTAL

Agenda

- Motivation: annotation schemes for cross-document coreference resolution (CDCR)
- Related work
- Dataset comparison: ECB+ vs NewsWCL50
  - Qualitative comparison
  - Quantitative comparison
- Discussion
- Conclusion

Motivation: annotation of cross-document coreference resolution datasets

Alexander Lukashenko locked and loaded to fight Belarus 'rats'

Tens of thousands of demonstrators massed in central Minsk on Sunday to demand the resignation of Belarusian President Alexander Lukashenko, who flew over the scene of the banned protest in a helicopter and called the marchers "rats".

The authoritarian leader, shown later clutching an assault rifle upon landing at his central Minsk residence, has ordered the military into full combat readiness in the face of the biggest challenge to his 26-year rule of the former Soviet socialist republic.

Source 1: The Australian

The Economist

Miffed in Minsk

Waving slippers at the "cockroach" president of Belarus

THEY CAME wielding slippers, with which to squish the man they call "the cockroach". Alexander Lukashenko, an idiosyncratic autocrat, has retained many of the oppressive structures and symbols of the Soviet Union in Belarus for the past quarter-century. But as the country heads towards a presidential election on August 9th, there are signs that the long-preserved edifice is crumbling.

Source 2: The Economist

Motivation: annotation of cross-document coreference resolution datasets

Alexander Lukashenko locked and loaded to fight Belarus 'rats'

Tens of thousands of demonstrators massed in central Minsk on Sunday to demand the resignation of Belarusian President Alexander Lukashenko, who flew over the scene of the banned protest in a helicopter and called the marchers "rats".

The authoritarian leader, shown later clutching an assault rifle upon landing at his central Minsk residence, has ordered the military into full combat readiness in the face of the biggest challenge to his 26-year rule of the former Soviet socialist republic.

Source 1: The Australian

The Economist

Miffed in Minsk

Waving slippers at the "cockroach" president of Belarus

THEY CAME wielding slippers, with which to squish the man they call "the cockroach". Alexander Lukashenko, an idiosyncratic autocrat, has retained many of the oppressive structures and symbols of the Soviet Union in Belarus for the past quarter-century. But as the country heads towards a presidential election on August 9th, there are signs that the long-preserved edifice is crumbling.

Source 2: The Economist

Research Question and Tasks

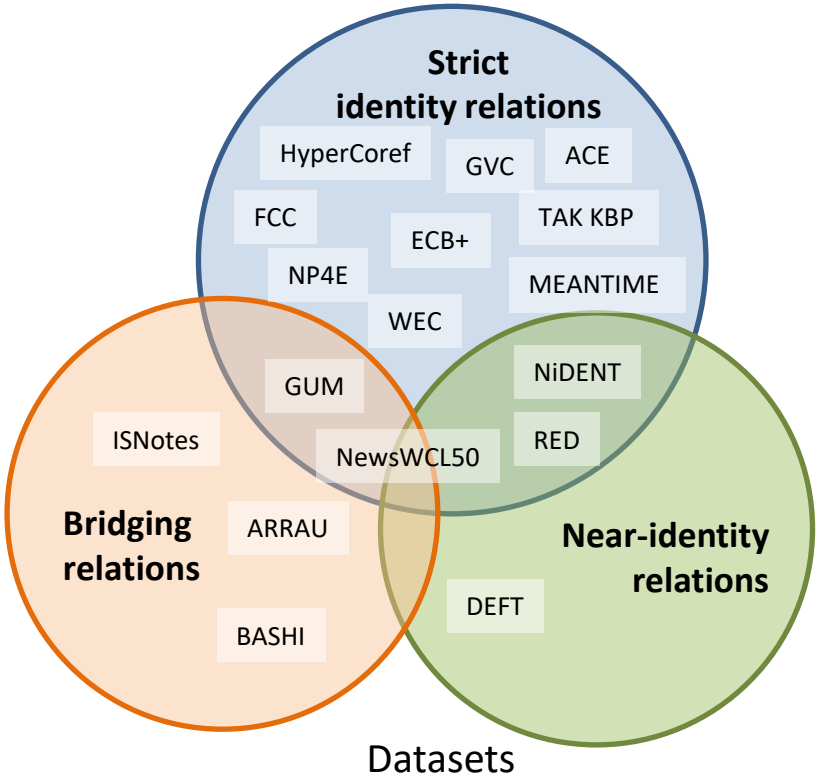
RQ: How to compare CDCR datasets and which aspects to consider in the analysis?

Research tasks

- Compare and discuss two perspectives of annotating coreference chains and relations:
  - **Event-centric** chains with **strict identity relations** (ECB+ [Cy14])
  - **Concept-centric** chains with loose coreference relations, i.e., **combination of identity and bridging anaphoric relations** (NewsWCL50 [Ha19])
- Discuss further directions of evaluation of the CDCR models

Related work

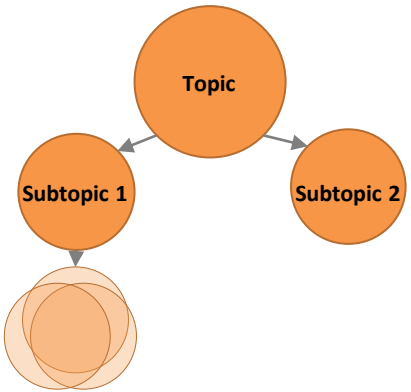
- CDCR datasets have two triggers for coreference chains
  - Event-centric (occurrence of an event matters)
  - Concept-centric (frequency of an event/entity matters)
- Typically separated research
  - Strict identity relations
  - Near-identity relations
  - Bridging relations
- Related work focuses on comparing (CD)CR datasets of same nature
  - For example, event-centric ECB+, FCC, GVC [Bu21]
- Very scarce comparison of the CDCR datasets of the different nature
  - We compare diverse ECB+ and NewsWCL50



Qualitative comparison

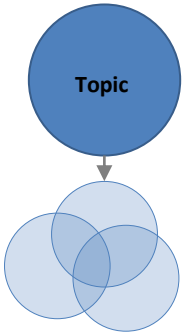
Dataset structure

- ECB+
  - *Topic*, e.g., earthquake
  - *Subtopic*, i.e., a narrowly-related event described by an action, participant(s), location, and time
  - *Article*, i.e., clearly represents a subtopic



Articles
Large content overlap

- NewsWCL50
  - *Topic*, i.e., articles reporting on the same event
  - *Article*, i.e., discusses at least partially different aspects of the reported event



Articles
Smaller content overlap, contain discussions of various aspects of the event

Annotation: Mentions & Relations

Dataset	Which mentions	How and which mentions to annotate	When to link into the coreference chains	Problems
ECB+	• action • participant • location • datetime (as attributes of an action)	• minimum-span style • NPs and VPs • pronouns	• entities: participant, location, datetime – if they refer to the same entity • events: same action + participant + location + datetime	• a few mentions of the entities that are not attributes of actions remain unannotated • many event-centric annotated mentions become singleton coreference chains
NewsWCL50	• action/event • actor • object • GPE • abstract entities • (no location) • (no datetime)	• maximum-span style • NPs and VPs • (no pronouns)	• minimum 5 occurrences of an event/entity linked with diverse relations across the related articles	• some entities are too abstract and need refinement • skipped annotations of location, datetime, and pronouns

Zhukova et al. "Towards Evaluation of Cross-document Coreference Resolution Models Using Datasets with Diverse Annotation Schemes"

Relations

ECB+: strict identity

- Actor, location, datetime: refer to the same entity
  "the U.S." – "America"
- Event components: same action, participant, location, datetime
  "Lindsay Lohan checked into *rehab*."
  "Ms. Lohan entered *a rehab facility*."

NewsWCL50: a mix of identity and bridging relations

(1) **identity relations** including **subject-predicate coreference**
"Donald Trump" – "the newly-minted president"
(2) **synonym relations**
"a caravan of immigrants" – "a few hundred asylum seekers"
(3) **metonymy**
"the Russian government" – "the Kremlin"
(4) **meronymy/holonymy**
"the U.S.-Mexico border" – "the San Ysidro port of entry"
(5) mentions are **linked with copular verbs**, e.g., "be," "seem," or "call"
"Trump called Kim Jong Un a Little Rocket Man" →
"Kim Jong Un" – "Little Rocket Man"
(6) mentions meet **GPE definitions as of ACE annotation**
"the U.S." – "American people"
(7) mentions are **elements of one set**
"guaranteed to disrupt trade" – "drove down steel prices" as members of a set "Consequences of tariff imposition"

Reannotation experiment

Head	ECB+ (original)	NewsWCL50 (following the coding book)
Jeffs	Jeffs x15 Warren Jeffs x1 polygamist sect leader Warren Jeffs x1 prophet Warren Jeffs x2 leader Warren Jeffs x2 polygamist Warren Jeffs x1 polygamist leader Warren Jeffs x3 Warren Jeffs, Polygamist Leader x1	Jeffs x66 Warren Jeffs x4 polygamist sect leader Warren Jeffs x2 prophet Warren Jeffs x1 FLDS prophet Warren Jeffs x1 Mr. Jeffs x2 the 54-year-old Jeffs x1 Jeffs, who acted as his own attorney x1 Jeffs, who was indicted more than two years ago x1 Warren Jeffs, leader of the Fundamentalist Church of Jesus Christ… x1 Warren Jeffs, polygamist leader x1
prophet	prophet x2	the prophet x1 the self-styled prophet x1 prophet of the Fundamentalist Church of the Jesus Christ… x1
head	head x1	the 55-year-old head of the Fundamentalist Church of Jesus Christ… x1 the ecclesiastical head of the Fundamentalist Church of Jesus Christ… x1
leader	leader x4 FLDS leader x1	their spiritual leader x1

Reannotation experiment

Head	ECB+	NewsWCL50
accomplice	--	an accomplice to rape by performing a marriage involving an underage girl x1 an accomplice to sexual conduct with a minor x1 an accomplice to the rape of a 14-year-old girl x1 an accomplice for his role x1
other	he, his, him, who x22 polygamist x2 attorney x2 pedophile x1	a victim of religious persecution x1 the defendant x1 the highest-profile defendant x1 president x1 the father of a 15-year-old FLSD member's child x1 her father x1 God's spokesman on earth x1 his client x1 their client x1 a problem x1 a handful from day one x1 this individual x1 one individual, Warren Steed Jeffs x1 one of the most wicked men on the face of the earth since the days of Father Adam x1 penitent x1

Quantitative comparison

Lexical diversity: previous metrics

- $F1_{CoNLL}$ of a simple same mentions' head-lemmas baseline [Cy14]
  - Evaluate combinations and links between annotated (key) and resolved (response) mentions
  - Average of F1: B^3, MUC, $CEAF_e$
  - High F1 = low lexical diversity
  - $F1_{CoNLL}$ is inflated with singletons [Ca21]

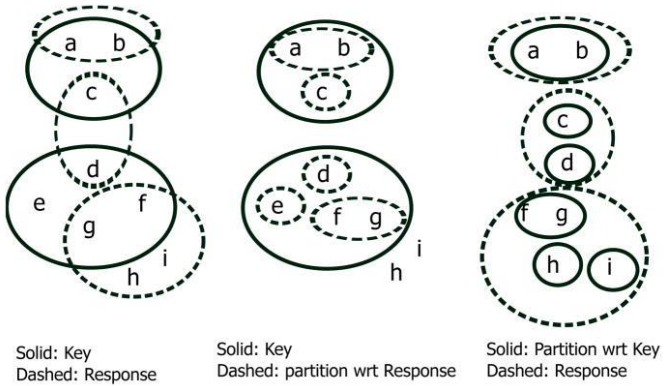
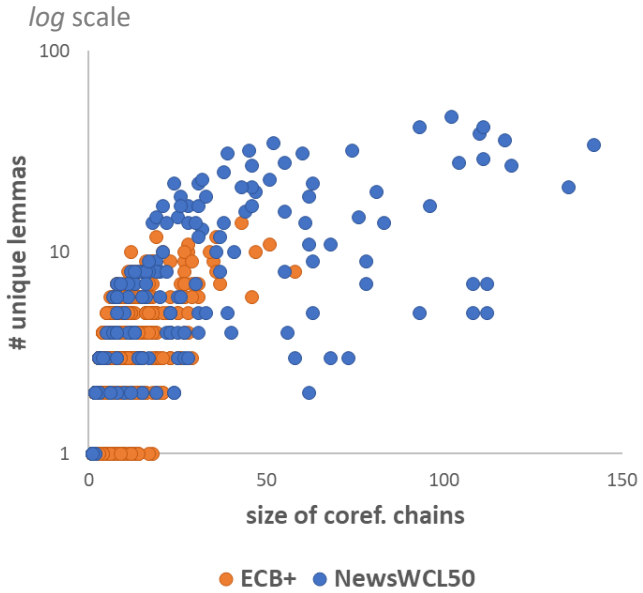


Figure source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667668/

- Unique number of mentions' head-lemmas [Ei21]
  - A quite discrete measure, no detailed comparison

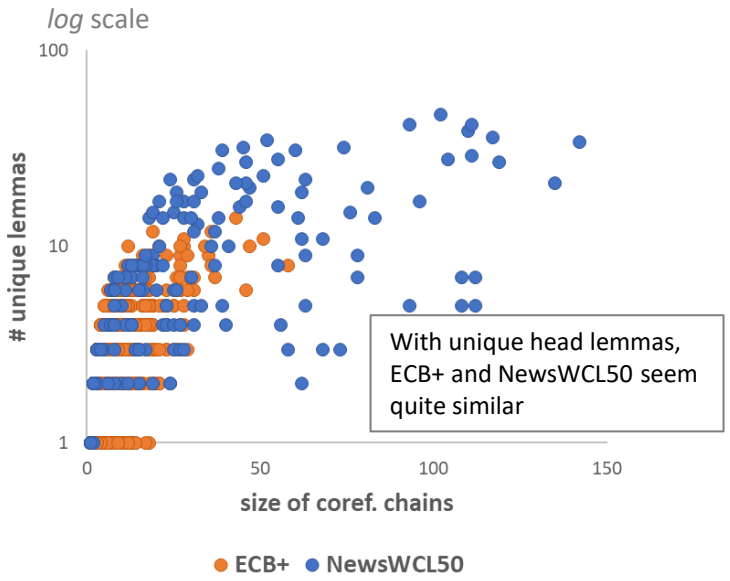


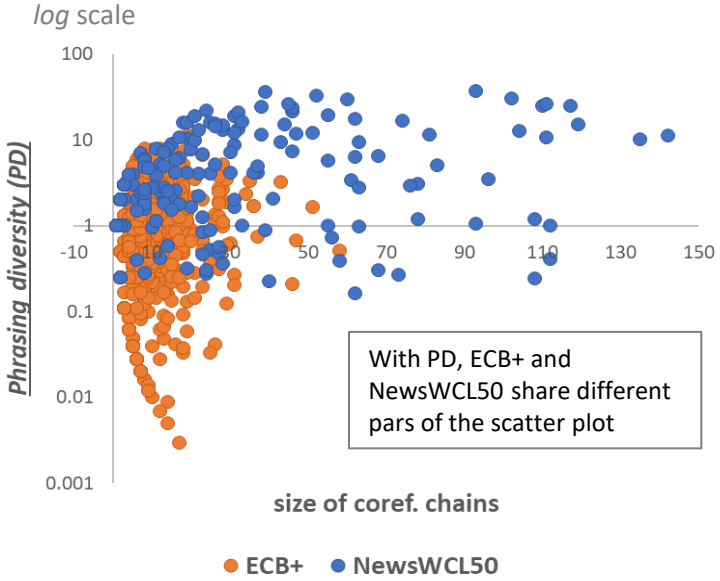
Zhukova et al. "Towards Evaluation of Cross-document Coreference Resolution Models Using Datasets with Diverse Annotation Schemes"

Lexical diversity: a new Phrasing diversity metric (PD)

Phrasing diversity (PD) measures lexical variation of coreference chains <u>using the mention frequency and phrasing variation.</u>

→ yields more detailed degrees the lexical diversity for the same coref.chains

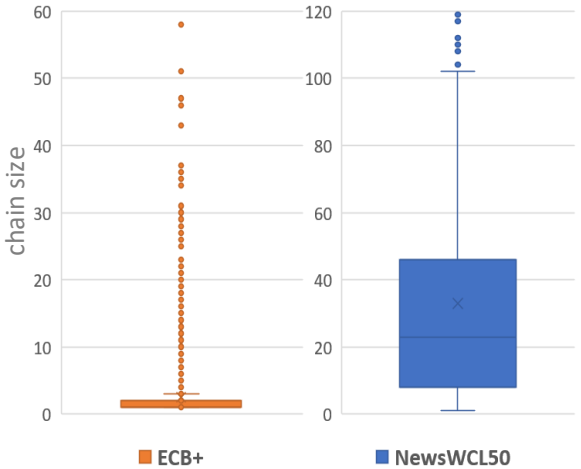




Numeric comparison

datasets	topics	subtopics	articles	tokens	mentions	event mentions	entity mentions	chains	singletons	inter-coder reliability (ICR)
ECB+	43	86	962	377 367	12 004	4 013	7 991	4 759	3 445	0.76
NewsWCL50	10	-	50	49 591	5600	1225	4375	170	10	0.65

	with singletons					without singletons				
datasets	average chain size	$F1_{CoNLL}$	average unique head lemmas	phrasing diversity (PD)	PD average	average chain size	$F1_{CoNLL}$	average unique head lemmas	phrasing diversity (PD)	PD average
ECB+	2.5	72.9	1.4	1.3	1.1	6.4	64.4	2.4	1.4	1.3
NewsWCL50	32.9	46.8	9.9	9.7	6.7	35	46.5	10.5	9.7	7.0



Coref.chain size distribution

Intercoder agreement

- ECB+: $\kappa = 0.74$ event/entity mention annotation, $\kappa = 0.76$ for chain annotation
- NewsWCL50: $AOA = 0.65$ (average observed agreement, less restrictive than κ)
  → more loosely related anaphora may result in lower ICR

Density of mention annotations per article

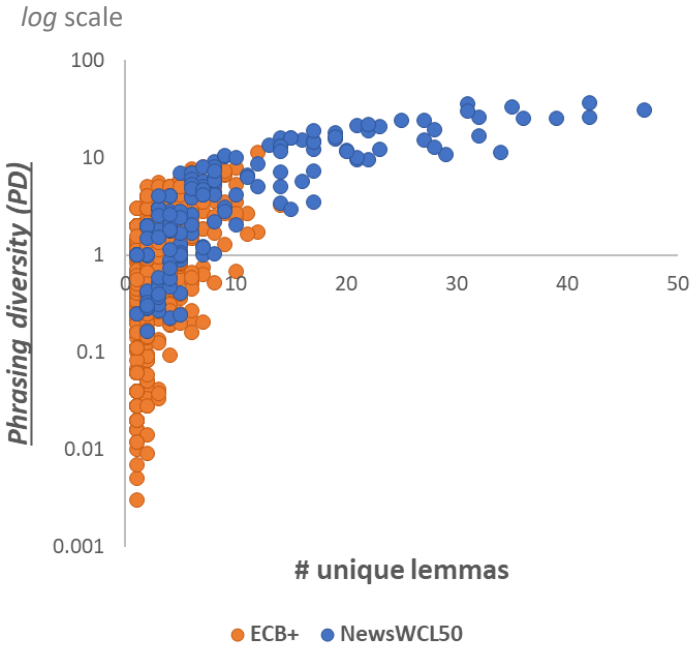
- ECB+: 13 per article
- NewsWCL50: 112 per article
  → annotation of NewsWCL50 aims at finding all occurrences of entities/events

Discussion

- The proposed PD metric measures lexical diversity in higher level of details compared to the unique head lemmas.
- ECB+
  - Subtopic level of events with similar wording
  - Only strict identity relation in coref.chains
  - Smaller coreference chains and lower lexical diversity

    → **lexical ambiguity problem** [Bu21]
- NewsWCL50
  - Topic level with large portion of new event aspects in each article
  - A mix of identity and more loose anaphoric relations → increased abstractness of the chains
  - Large coreference chains and high lexical diversity

    → **lexical diversity problem**



CDCR models need to be evaluated on a diverse set of CDCR datasets with different focus and challenges.

Conclusion

Conclusion

- Compared and discussed two diverse CDCR datasets
  - Event-centric ECB+ with *strict identity coreference relations*
  - Concept-centric NewsWCL50 with *mixed identity and bridging coreference relations*
- Bugert et al. 2021 evaluated the robustness of the CDCR models to handle multiple event-centric datasets → **a lexical disambiguation challenge**
- New phrasing diversity metric (PD): represents frequency and lexical variation of the mentions in coreference chains.
- $PD_{ECB+} < PD_{NewsWCL50}$ → a need for evaluation on **a lexical diversity challenge**
- Lexical disambiguation challenge + lexical diversity challenge

  → **generalizability and robustness of the CDCR models**

References

- **[Bu21]** Bugert, M., Reimers, N., and Gurevych, I. (2021). Generalizing cross-document event coreference resolution across multiple corpora. Computational Linguistics, pages 1–43.
- **[Ca21]** Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., and Dagan, I. (2021). Realistic evaluation principles for cross-document coreference resolution. In Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, pages 143–151, Online, August. Association for Computational Linguistics.
- **[Cy14]** Cybulska, A. and Vossen, P. (2014). Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 4545–4552, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
- **[Ei21]** Eirew, A., Cattan, A., and Dagan, I. (2021). WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2498–2510, Online, June. Association for Computational Linguistics.
- **[Ha19]** Hamborg, F., Zhukova, A., and Gipp, B. (2019). Automated identification of media bias by word choice and labeling in news articles. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL).

Questions

Anastasia Zhukova
Data & Knowledge Engineering Group
zhukova@uni-wuppertal.de
@ana_m_zhukova

Comparison of ECB+ and NewsWCL50:

- Parsed into same formats
- Contain scripts for the reported evaluation metrics

https://github.com/anastasia-zhukova/Diverse_CDCR_datasets

Backup slides

Phrasing diversity metric (PD) metric

On the chain level

$$PD_c = \begin{cases} \frac{\sum_{\forall h \in H_c} \frac{|U_h|}{|P_h|} \cdot \sum_{\forall h \in H_c} |U_h|}{|M_c|}, & \text{if } |M_c| > 1 \\ 1, & \text{else} \end{cases}$$

where H_c represents all unique heads of phrases of an annotated chain c,
$|U_h|$ is the number of unique phrases with h,
$|P_h|$ is the number of all phrases with head h,
$|M_c|$ is a number of all mentions of an annotated chain c.

Weighted-average on a topic level

$$PD = \frac{\sum_{\forall c \in C} |M_c| \cdot PD_c}{\sum_{\forall c \in C} |M_c|}$$

Zhukova et al. "Towards Evaluation of Cross-document Coreference Resolution Models Using Datasets with Diverse Annotation Schemes"