My Intended Dissertation
It's (Retrospective) Disruption
And My Current Research

Yanai Elazar

Georgetown University, 19th April, 2023

Hi There

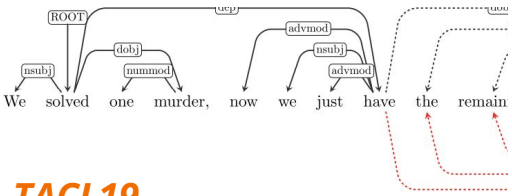
Yanai Elazar, Postdoc @ AI2 & UW

My Intended Dissertation: Missing Elements

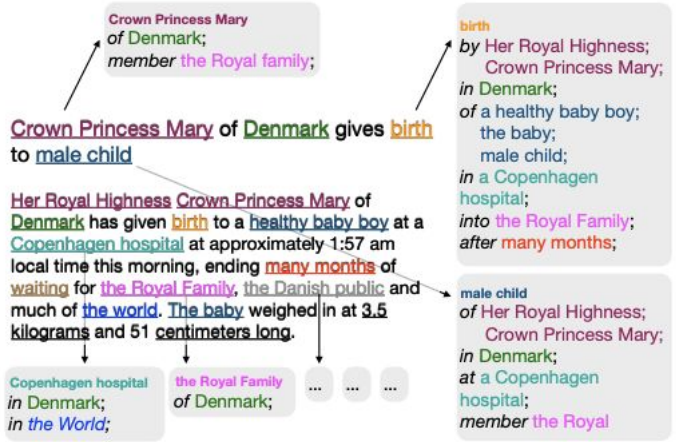
Where
for Nun

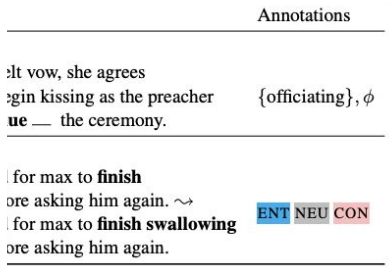
Text-based NP Enrichment

Yanai Elazar* Victoria Basmov* Yoav Goldberg Reut Tsarfaty
Computer Science Department, Bar Ilan University
{yanaiela,vikasa arfaty}@gmail.com



TACL19





TACL22

My Intended Dissertation: Missing Elements

- Language constructions with implicit information
- Spans across different constructions
  - Syntactic (fused heads: “I bought 5 apples but got only **4**”)
  - Semantics (complement coercion: “I **started** a book”)
  - Pragmatics/Semantic/Commonsense (TNE: “I entered the room and the **windows** were open”)
- The motivation:
  - Text understanding
  - Text augmentation

Text-based NP Enrichment

Yanai Elazar*, Victoria Basmov*,
Yoav Goldberg, Reut Tsarfaty

TACL 2022

TNE: a New Benchmark for Reading Comprehension

- Extracting relations between NPs
- A useful task
- A challenging benchmark

TNE: a New Benchmark for Reading Comprehension

Church of Scientology does not see humor in website dedicated to Tom Cruise

September 25, 2005

http://ScienTOMogy.info has apparently received a fax and at least 6 emails in the span of 2 days from Scientology lawyer Ava Paquette of Moxon & Kobrin threatening **a lawsuit** of up to $100,000 if the domain name ownership is not transferred. This type of letter is often called a cease and desist letter.

The owners of scienTOMogy.info have posted the complaints and their replies, saying that the site simply expresses opinion, does not make any claims, and clearly states that it has no connection to the Church of Scientology. "The site was put up as a single source to view all the recent hype Tom has made about the church - it does nothing but show Tom, so we are at a loss as to why the church is acting so rashly."

The Church of Scientology is notorious for pursing legal action against its critics, under the name of the "Religious Technology Center" (RTC). It previously made headlines when it used the US's Digital Millennium Copyright Act to remove xenu.net, a site critical of Scientology, from Google's listings.

NPs:

Church of Scientology | humor | website | Tom Cruise
September 25, 2005 | a fax | at least 6 emails | the span of 2 days
Scientology lawyer Ava Paquette | Moxon & Kobrin | a lawsuit
the domain name ownership | This type | letter | The owners
the complaints | their replies | the site | opinion | any claims
no connection | the Church of Scientology | The site | a single source
all the recent hype | Tom | the church | nothing | Tom | we | a loss
the church | The Church of Scientology | legal action | its critics
the name | the "Religious Technology Center" | RTC | headlines
the US's Digital Millennium Copyright Act | xenu.net | a site | Scientology

Coreference Chains:

1. Church of Scientology the Church of Scientology the church the church The Church of Scientology
2. humor
3. website the site The site a single source
4. Tom Cruise Tom Tom
5. a fax
6. at least 6 emails
7. the span of 2 days
8. Scientology lawyer Ava Paquette
9. Moxon & Kobrin
10. a lawsuit
11. the domain name ownership
12. This type

NP Relations:

1. a lawsuit [about the complaints]
2. a lawsuit [against The owners we]
3. a lawsuit [about the domain name ownership]
4. a lawsuit [from Moxon & Kobrin]
5. a lawsuit [from Scientology lawyer Ava Paquette]
6. a lawsuit [against website the site The site a single source]
7. a lawsuit [by Church of Scientology the Church of Scientology the church the church The Church of Scientology]
8. a lawsuit [in US]

TNE: a New Benchmark for Reading Comprehension

-
-
-

Church of Scientology does not see humor in website dedicated to Tom Cruise

September 25, 2005
http://ScienTOMogy.info has apparently received a fax and at least 6 emails in the span of 2 days from Scientology lawyer Ava Paquette of Moxon & Kobrin threatening **a lawsuit** of up to $100,000 if the domain name ownership is not transferred. This type of letter is often called a cease and desist letter.

The owners of scienTOMogy.info have posted the complaints and their replies, saying that the site simply expresses opinion, does not make any claims, and clearly states that it has no connection to the Church of Scientology. "The site was put up as a single source to view all the recent hype Tom has made about the church - it does nothing but show Tom, so we are at a loss as to why the church is acting so rashly."

The Church of Scientology is notorious for pursing legal action against its critics, under the name of the "Religious Technology Center" (RTC). It previously made headlines when it used the US's Digital Millennium Copyright Act to remove xenu.net, a site critical of Scientology, from Google's listings.

NPs:

Church of Scientology | humor | website | Tom Cruise
September 25, 2005 | a fax | at least 6 emails | the span of 2 days
Scientology lawyer Ava Paquette | Moxon & Kobrin | a lawsuit
the domain name ownership | This type | letter | The owners
the complaints | their replies | the site | opinion | any claims
no connection | the Church of Scientology | The site | a single source
all the recent hype | Tom | the church | nothing | Tom | we | a loss
the church | The Church of Scientology | legal action | its critics
the name | the "Religious Technology Center" | RTC | headlines
the US's Digital Millennium Copyright Act | xenu.net | a site | Scientology

Coreference Chains:

1. Church of Scientology the Church of Scientology the church the church The Church of Scientology
2. humor
3. website the site The site a single source
4. Tom Cruise Tom Tom
5. a fax
6. at least 6 emails
7. the span of 2 days
8. Scientology lawyer Ava Paquette
9. Moxon & Kobrin
10. a lawsuit
11. the domain name ownership
12. This type

NP Relations:

1. a lawsuit [about the complaints]
2. a lawsuit [against The owners we]
3. a lawsuit [about the domain name ownership]
4. a lawsuit [from Moxon & Kobrin]
5. a lawsuit [from Scientology lawyer Ava Paquette]
6. a lawsuit [against website the site The site a single source]
7. a lawsuit [by Church of Scientology the Church of Scientology the church the church The Church of Scientology]
8. a lawsuit [in US]

TNE: a New Benchmark for Reading Comprehension

Church of Scientology does not see humor in website dedicated to Tom Cruise

September 25, 2005

http://ScienTOMogy.info has apparently received a fax and at least 6 emails in the span of 2 days from Scientology lawyer Ava Paquette of Moxon & Kobrin threatening **a lawsuit** of up to $100,000 if the domain name ownership is not transferred. This type of letter is often called a cease and desist letter.

The owners of scienTOMogy.info have posted the complaints and their replies, saying that the site simply expresses opinion, does not make any claims, and clearly states that it has no connection to the Church of Scientology. "The site was put up as a single source to view all the recent hype Tom has made about the church - it does nothing but show Tom, so we are at a loss as to why the church is acting so rashly."

The Church of Scientology is notorious for pursing legal action against its critics, under the name of the "Religious Technology Center" (RTC). It previously made headlines when it used the US's Digital Millennium Copyright Act to remove xenu.net, a site critical of Scientology, from Google's listings.

NPs:

Church of Scientology | humor | website | Tom Cruise
September 25, 2005 | a fax | at least 6 emails | the span of 2 days
Scientology lawyer Ava Paquette | Moxon & Kobrin | a lawsuit
the domain name ownership | This type | letter | The owners
the complaints | their replies | the site | opinion | any claims
no connection | the Church of Scientology | The site | a single source
all the recent hype | Tom | the church | nothing | Tom | we | a loss
the church | The Church of Scientology | legal action | its critics
the name | the "Religious Technology Center" | RTC | headlines
the US's Digital Millennium Copyright Act | xenu.net | a site | Scientology

Coreference Chains:

1. Church of Scientology the Church of Scientology the church the church The Church of Scientology
2. humor
3. website the site The site a single source
4. Tom Cruise Tom Tom
5. a fax
6. at least 6 emails
7. the span of 2 days
8. Scientology lawyer Ava Paquette
9. Moxon & Kobrin
10. a lawsuit
11. the domain name ownership
12. This type

NP Relations:

1. a lawsuit [about the complaints]
2. a lawsuit [against The owners we]
3. a lawsuit [about the domain name ownership]
4. a lawsuit [from Moxon & Kobrin]
5. a lawsuit [from Scientology lawyer Ava Paquette]
6. a lawsuit [against website the site The site a single source]
7. a lawsuit [by Church of Scientology the Church of Scientology the church the church The Church of Scientology]
8. a lawsuit [in US]

TNE: a New Benchmark for Reading Comprehension

Church of Scientology does not see humor in website dedicated to Tom Cruise

September 25, 2005

http://ScienTOMogy.info has apparently received a fax and at least 6 emails in the span of 2 days from Scientology lawyer Ava Paquette of Moxon & Kobrin threatening **a lawsuit** of up to $100,000 if the domain name ownership is not transferred. This type of letter is often called a cease and desist letter.

The owners of scienTOMogy.info have posted the complaints and their replies, saying that the site simply expresses opinion, does not make any claims, and clearly states that it has no connection to the Church of Scientology. "The site was put up as a single source to view all the recent hype Tom has made about the church - it does nothing but show Tom, so we are at a loss as to why the church is acting so rashly."

The Church of Scientology is notorious for pursing legal action against its critics, under the name of the "Religious Technology Center" (RTC). It previously made headlines when it used the US's Digital Millennium Copyright Act to remove xenu.net, a site critical of Scientology, from Google's listings.

NPs:

Church of Scientology | humor | website | Tom Cruise
September 25, 2005 | a fax | at least 6 emails | the span of 2 days
Scientology lawyer Ava Paquette | Moxon & Kobrin | a lawsuit
the domain name ownership | This type | letter | The owners
the complaints | their replies | the site | opinion | any claims
no connection | the Church of Scientology | The site | a single source
all the recent hype | Tom | the church | nothing | Tom | we | a loss
the church | The Church of Scientology | legal action | its critics
the name | the "Religious Technology Center" | RTC | headlines
the US's Digital Millennium Copyright Act | xenu.net | a site | Scientology

Coreference Chains:

1. Church of Scientology the Church of Scientology the church the church The Church of Scientology
2. humor
3. website the site The site a single source
4. Tom Cruise Tom Tom
5. a fax
6. at least 6 emails
7. the span of 2 days
8. Scientology lawyer Ava Paquette
9. Moxon & Kobrin
10. a lawsuit
11. the domain name ownership
12. This type

NP Relations:

1. a lawsuit [about the complaints]
2. a lawsuit [against The owners we]
3. a lawsuit [about the domain name ownership]
4. a lawsuit [from Moxon & Kobrin]
5. a lawsuit [from Scientology lawyer Ava Paquette]
6. a lawsuit [against website the site The site a single source]
7. a lawsuit [by Church of Scientology the Church of Scientology the church the church The Church of Scientology]
8. a lawsuit [in US]

TNE: a New Benchmark for Reading Comprehension

Church of Scientology does not see humor in website dedicated to Tom Cruise

September 25, 2005

http://ScienTOMogy.info has apparently received a fax and at least 6 emails in the span of 2 days from Scientology lawyer Ava Paquette of Moxon & Kobrin threatening **a lawsuit** of up to $100,000 if the domain name ownership is not transferred. This type of letter is often called a cease and desist letter.

The owners of scienTOMogy.info have posted the complaints and their replies, saying that the site simply expresses opinion, does not make any claims, and clearly states that it has no connection to the Church of Scientology. "The site was put up as a single source to view all the recent hype Tom has made about the church - it does nothing but show Tom, so we are at a loss as to why the church is acting so rashly."

The Church of Scientology is notorious for pursing legal action against its critics, under the name of the "Religious Technology Center" (RTC). It previously made headlines when it used the US's Digital Millennium Copyright Act to remove xenu.net, a site critical of Scientology, from Google's listings.

NPs:

Church of Scientology | humor | website | Tom Cruise
September 25, 2005 | a fax | at least 6 emails | the span of 2 days
Scientology lawyer Ava Paquette | Moxon & Kobrin | a lawsuit
the domain name ownership | This type | letter | The owners
the complaints | their replies | the site | opinion | any claims
no connection | the Church of Scientology | The site | a single source
all the recent hype | Tom | the church | nothing | Tom | we | a loss
the church | The Church of Scientology | legal action | its critics
the name | the "Religious Technology Center" | RTC | headlines
the US's Digital Millennium Copyright Act | xenu.net | a site | Scientology

Coreference Chains:

1. Church of Scientology the Church of Scientology the church the church The Church of Scientology
2. humor
3. website the site The site a single source
4. Tom Cruise Tom Tom
5. a fax
6. at least 6 emails
7. the span of 2 days
8. Scientology lawyer Ava Paquette
9. Moxon & Kobrin
10. a lawsuit
11. the domain name ownership
12. This type

NP Relations:

1. a lawsuit [about the complaints]
2. a lawsuit [against The owners we]
3. a lawsuit [about the domain name ownership]
4. a lawsuit [from Moxon & Kobrin]
5. a lawsuit [from Scientology lawyer Ava Paquette]
6. a lawsuit [against website the site The site a single source]
7. a lawsuit [by Church of Scientology the Church of Scientology the church the church The Church of Scientology]
8. a lawsuit [in US]

Reading Comprehension

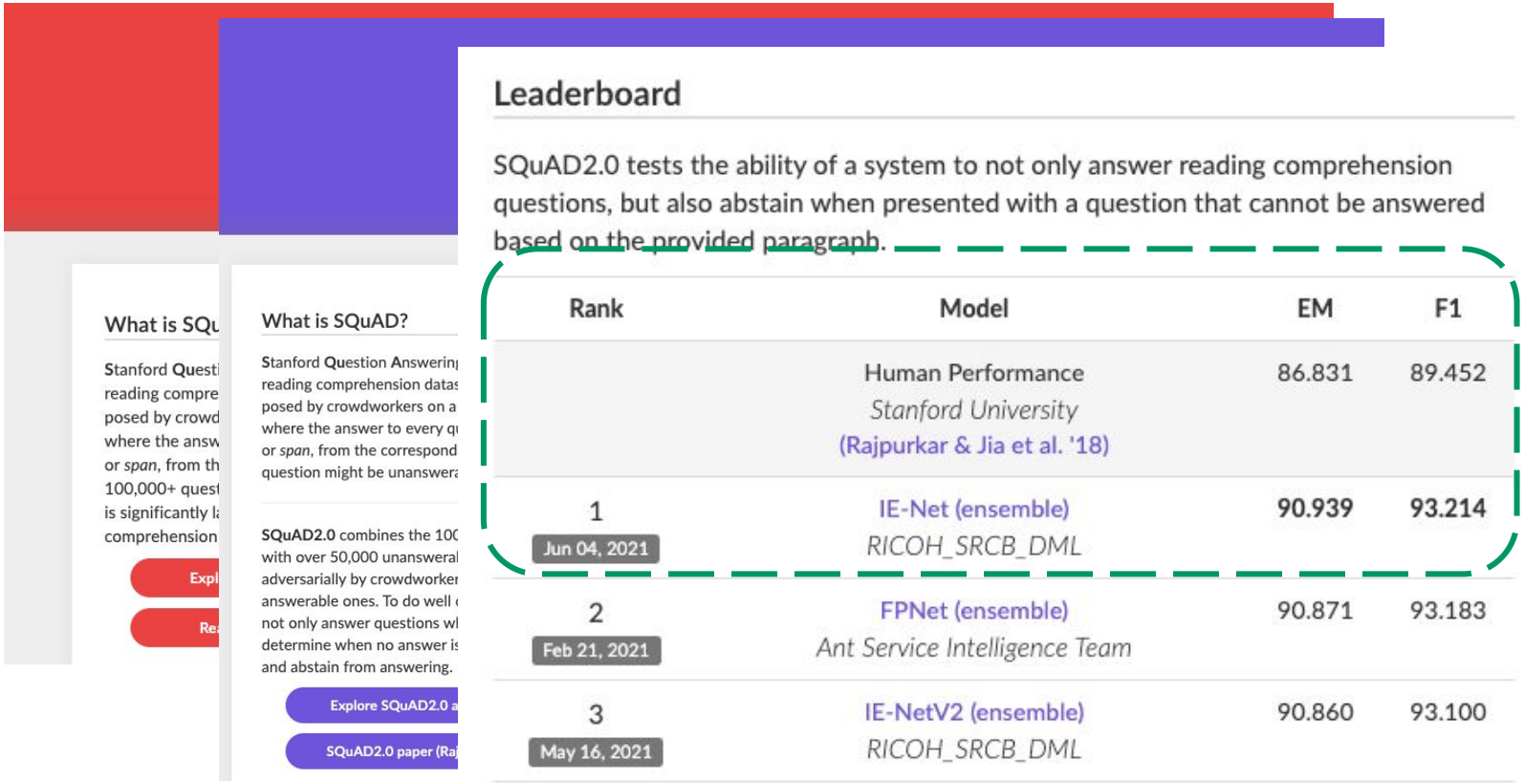
Towards a Reading Comprehension Benchmark

Today:

Reading Comprehension ≅ Question Answering

Towards a Reading Comprehension Benchmark

Today:



What is SQu

Stanford Questi
reading compre
posed by crowd
where the answ
or *span*, from th
100,000+ quest
is significantly l
comprehension

Expl

Re

What is SQuAD?

Stanford Question Answering
reading comprehension datas
posed by crowdworkers on a
where the answer to every q
or *span*, from the correspond
question might be unanswera

SQuAD2.0 combines the 100
with over 50,000 unanswera
adversarially by crowdworker
answerable ones. To do well
not only answer questions wh
determine when no answer is
and abstain from answering.

Explore SQuAD2.0 a

SQuAD2.0 paper (Ra

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100

Towards a Reading Comprehension Benchmark

Today:

Are we there yet???

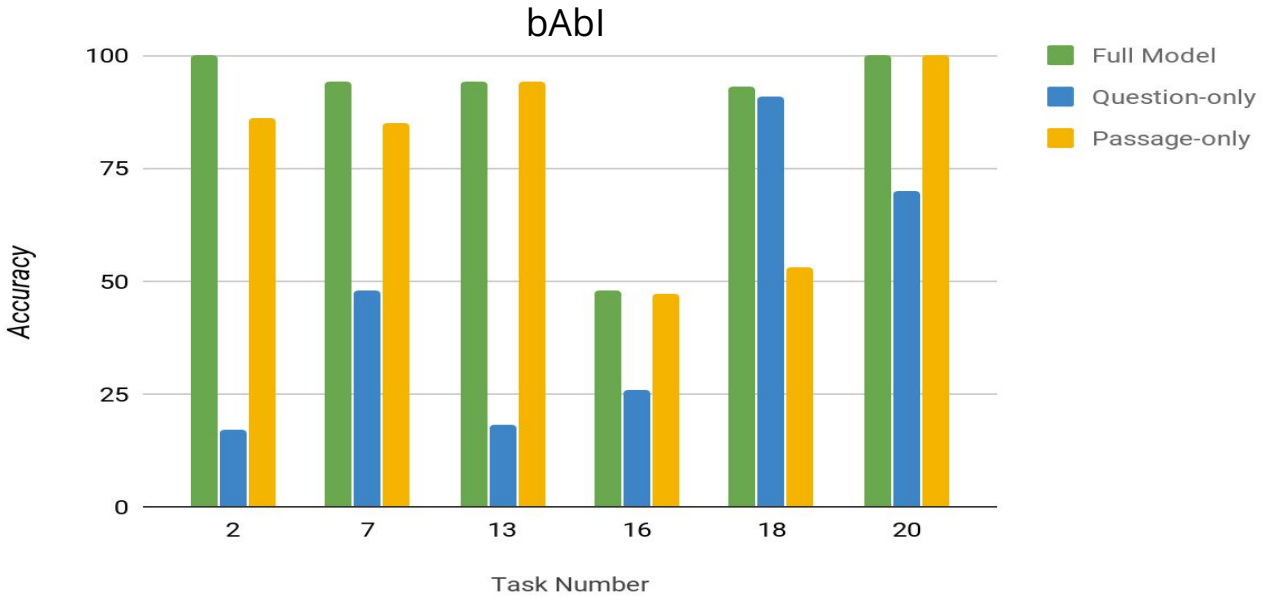
Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100

Towards a Reading Comprehension Benchmark

Today:



Are we there yet???
NO!

Kaushik & Lipton, 2018
Thanks Divyansh for the graph!

Towards a Reading Comprehension Benchmark

- A big drawback of QA as an RC benchmark is the modeling:

RC-Eval(Text, Question)

Instead of:

RC-Eval(Text)

Towards a Reading Comprehension Benchmark

English Machine Reading Comprehension Datasets: A Survey

Abstract

This paper surveys 60 English Machine Reading Comprehension datasets, with a view to providing a convenient resource for other researchers interested in this problem. We categorize the datasets according to their question and answer form and compare them across various dimensions including size, vocabulary, data source, method of creation, human performance level, and first question word. Our analysis reveals that Wikipedia is by far the most common data source and that there is a relative lack of *why*, *when*, and *where* questions across datasets.

The community really cares about Reading Comprehension

But... most of the benchmarks revolve around Question Answering

Towards a Reading Comprehension Benchmark

But what does it mean for machines to comprehend texts?

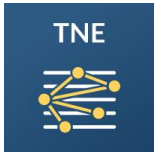
To Test Machine Comprehension, Start by Defining Comprehension

"... First, we argue that existing approaches do not adequately define comprehension; they are too unsystematic about what content is tested..."

Towards a Reading Comprehension Benchmark

Can we do better?

Text-based NP Enrichment: (TNE)



TNE: An Alternative to Reading Comprehension

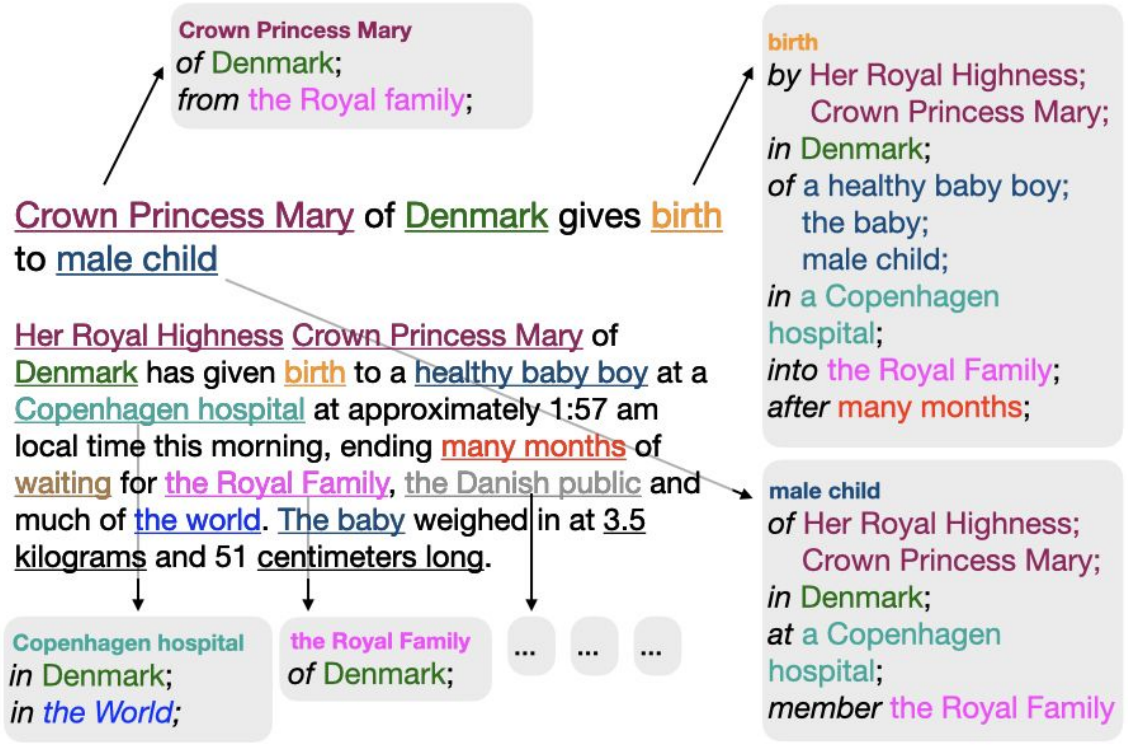
and a useful task...

TNE

We propose TNE, as an alternative benchmark for measuring RC

- We model (and benchmark) only the text (no immediate artifacts)
- We have a well-scoped task...
  that focuses on the relations between NPs in a (longish) text
- It is a useful task

TNE

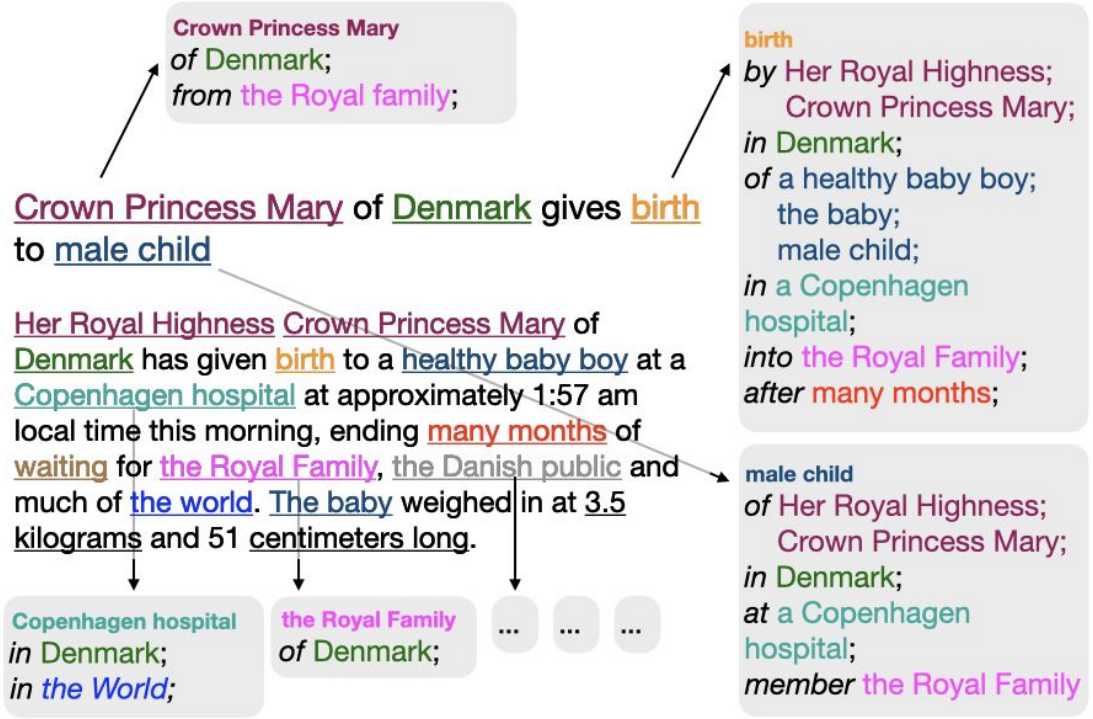

Crown Princess Mary
of Denmark;
from the Royal family;
birth
by Her Royal Highness;
Crown Princess Mary;
in Denmark;
of a healthy baby boy;
the baby;
male child;
in a Copenhagen
hospital;
into the Royal Family;
after many months;
Crown Princess Mary of Denmark gives birth
to male child
Her Royal Highness Crown Princess Mary of
Denmark has given birth to a healthy baby boy at a
Copenhagen hospital at approximately 1:57 am
local time this morning, ending many months of
waiting for the Royal Family, the Danish public and
much of the world. The baby weighed in at 3.5
kilograms and 51 centimeters long.
male child
of Her Royal Highness;
Crown Princess Mary;
in Denmark;
at a Copenhagen
hospital;
member the Royal Family
Copenhagen hospital
in Denmark;
in the World;
the Royal Family
of Denmark;
...
...
...

TNE: The Task

- Input:

TNE: The Task: Relations between NPs

- Output:



Relation between NPs

Have been studied before in NLP…

- **SRL:** relations mediated via verbs

<Crown Princess Mary, **gives**, birth>

Her Royal Highness Crown Princess Mary of Denmark has given birth to a healthy baby boy at a Copenhagen hospital at approximately 1:57 am local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

Relation between NPs

Have been studied before in NLP...

- **SRL:** relations mediated via verbs
- **Coreference Resolution:** is same as relation

<Crown Princess Mary, **identity**, Her Royal Highness Crown Princess Mary>

Relation between NPs

<male child, **of** Crown Princess Mary>

Have been studied before in NLP…

- **SRL:** relations mediated via verbs
- **Coreference Resolution:** is same as relation
- **Bridging:** (a limited set of) preposition mediated relations*

* not only

Relation between NPs

<male child, **of** Crown Princess Mary>

TNE

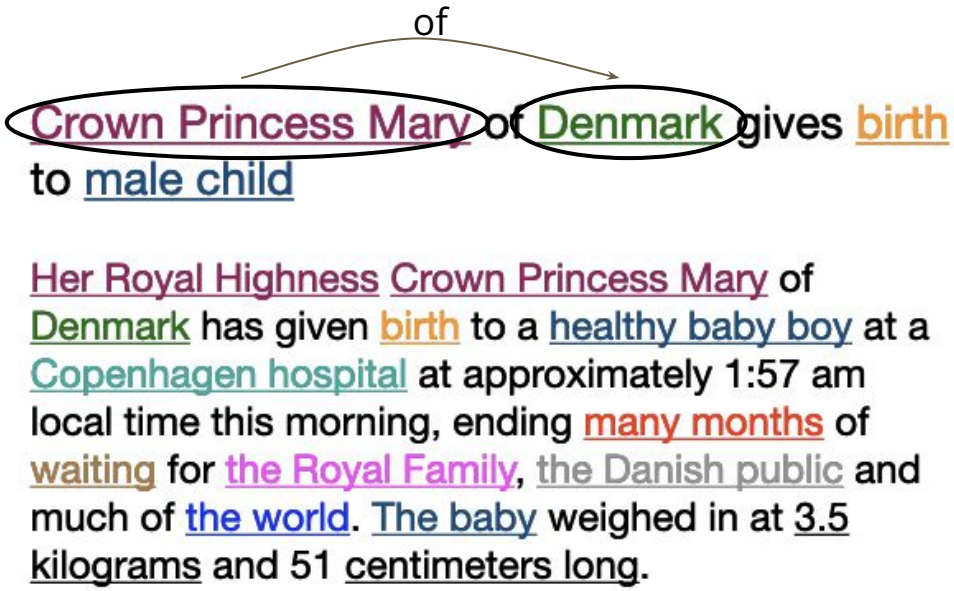
- Collect all preposition-mediated relations

Relation between NPs

<Crown Princess Mary, **of**, Denmark>

TNE

- Collect all preposition-mediated relations
- Explicit (rare)

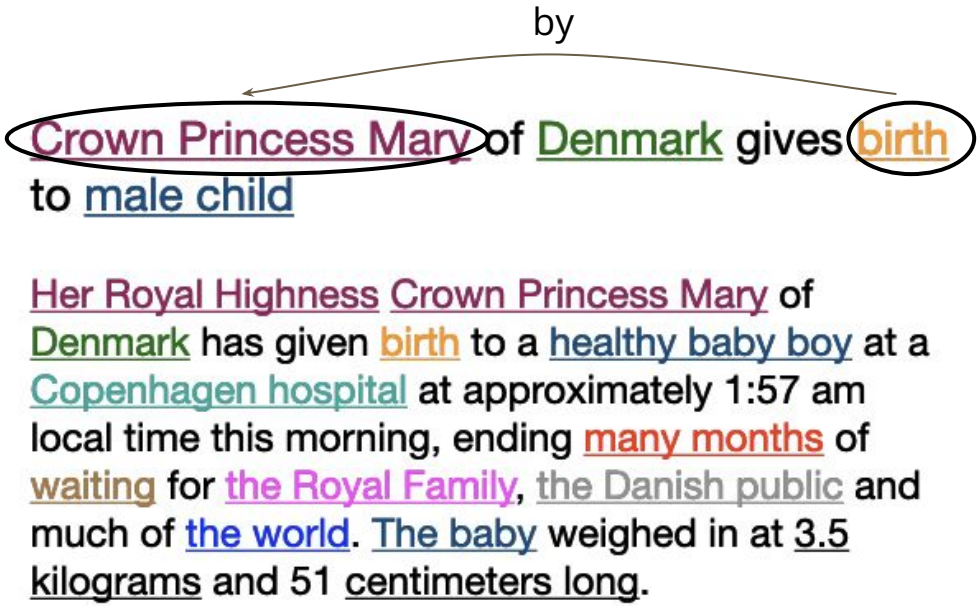


Relation between NPs

TNE

- Collect all preposition-mediated relations
- Explicit (rare)
- Implicit (common)

<birth, **by**, Crown Princess Mary>

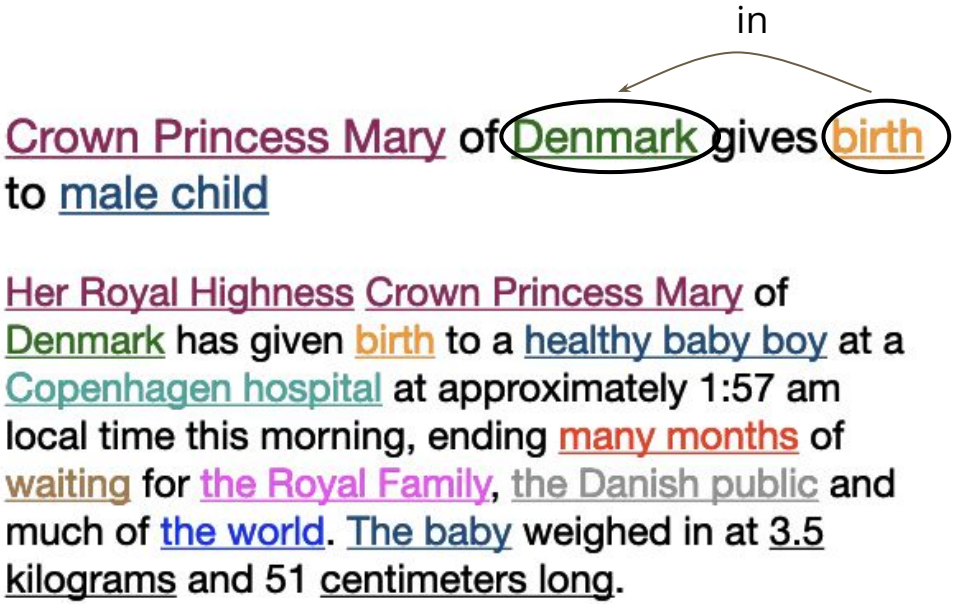


Relation between NPs

TNE

- Collect all preposition-mediated relations
- Explicit (rare)
- Implicit (common)

<birth, **in**, Denmark>

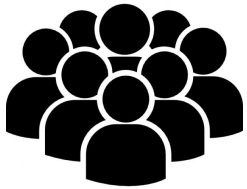


TNE: The Dataset

- Collecting using **Controlled Crowdsourcing** (*Roit et al., 20, Pyatkin et al. 20*) with a (very!) carefully crafted collection design

TNE: The Dataset

- Collecting using **Controlled Crowdsourcing** (*Roit et al., 20, Pyatkin et al. 20*) with a (very!) carefully crafted collection design
- 23 Annotators



TNE: The Dataset

- Collecting using **Controlled Crowdsourcing** (*Roit et al., 20, Pyatkin et al. 20*) with a (very!) carefully crafted collection design
- 23 Annotators
- 5.5+$ per document

TNE: The Dataset

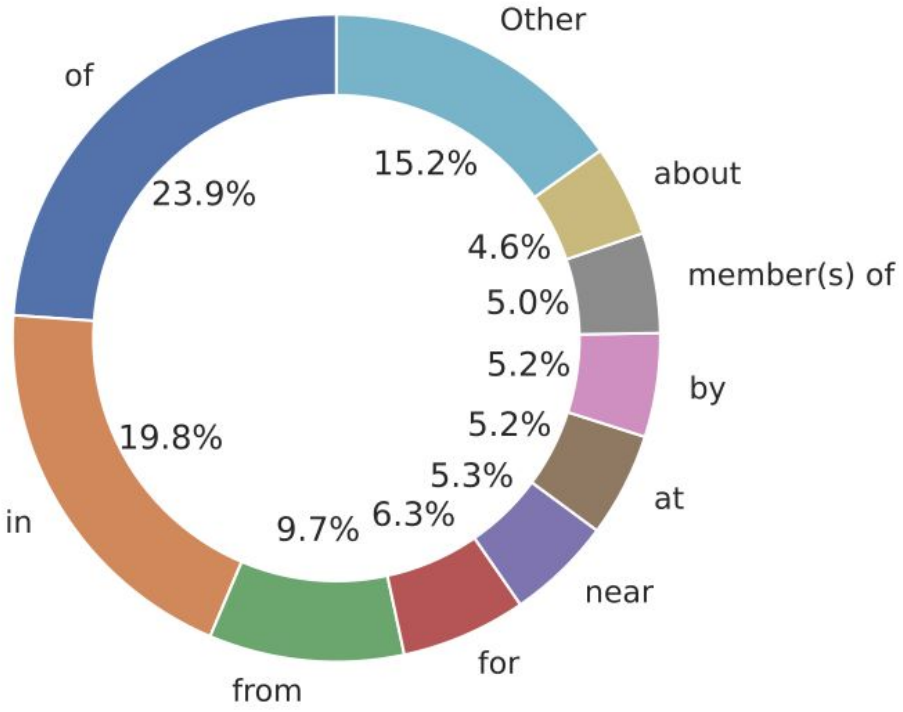
- Collecting using **Controlled Crowdsourcing** (*Roit et al., 20, Pyatkin et al. 20*) with a (very!) carefully crafted collection design
- 23 Annotators
- 5.5+$ per document
- High agreement scores! 88.9-94.4 F1

TNE: The Dataset

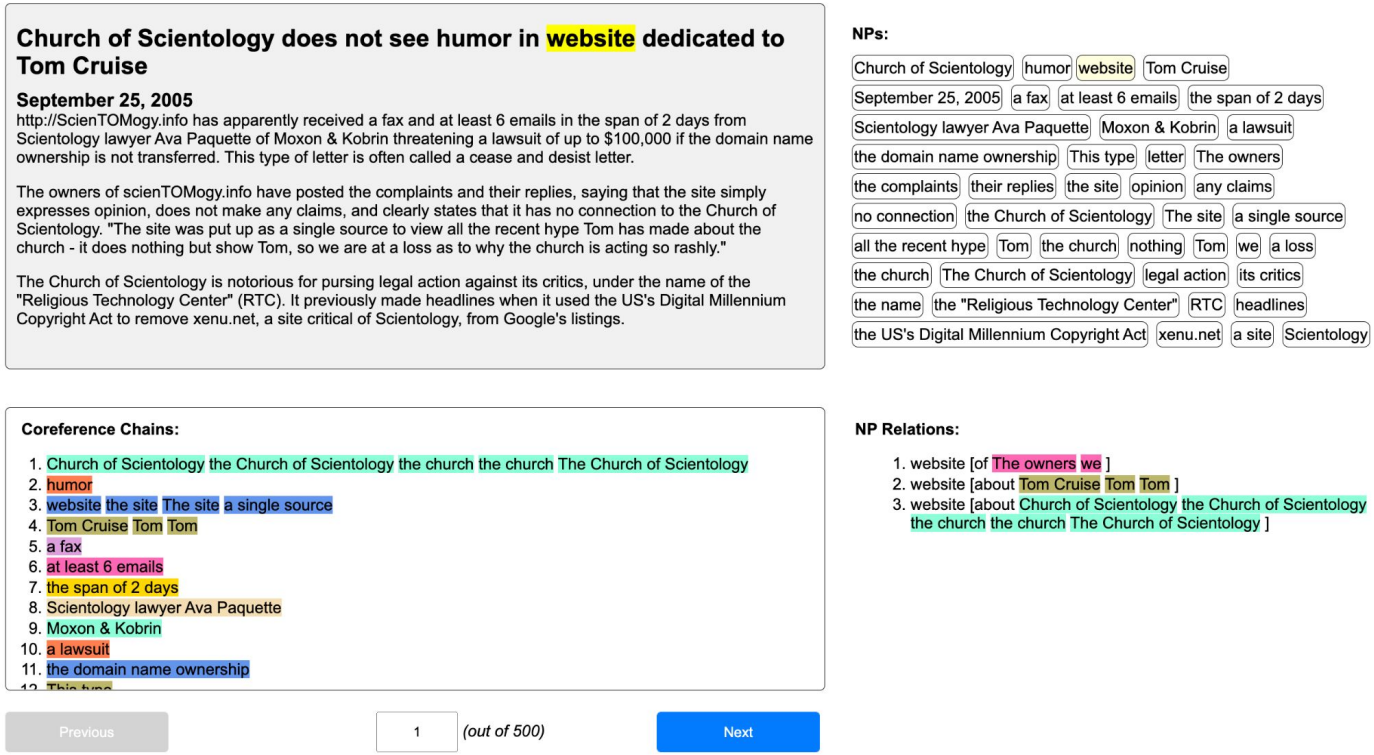
- 5.5K (English) Documents:
  - 4000 for train
  - 500 for dev
  - 500 for test
  - 500 for OOD-test
- Title + 3 paragraphs per document
- **1,000,000+** Links in total in the dataset
- 186 NP links per document (on average)
  Out of 36^2-36=**1260 possible links** per document (~15%)

TNE: The Dataset

- 23 prepositions



TNE: The Dataset



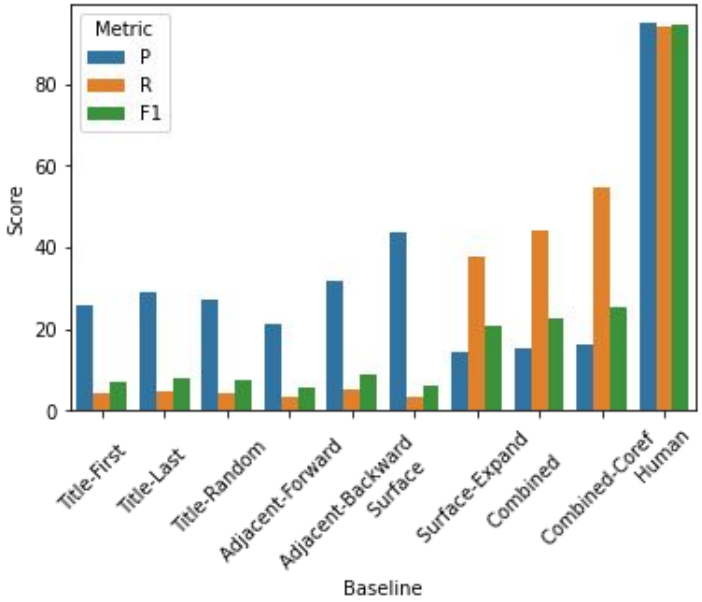
TNE: Evaluation & Models

TNE: Evaluation

- F1 score of the recovered triplets (NP1, preposition, NP2) between the predictions and the gold-standard
- Precision and recall as well
- *Labeled* and *Unlabaled* versions (in this talk I only report labeled)

TNE: Heuristic Baselines

- We live in a world of datasets biases
- Do we have any of these?



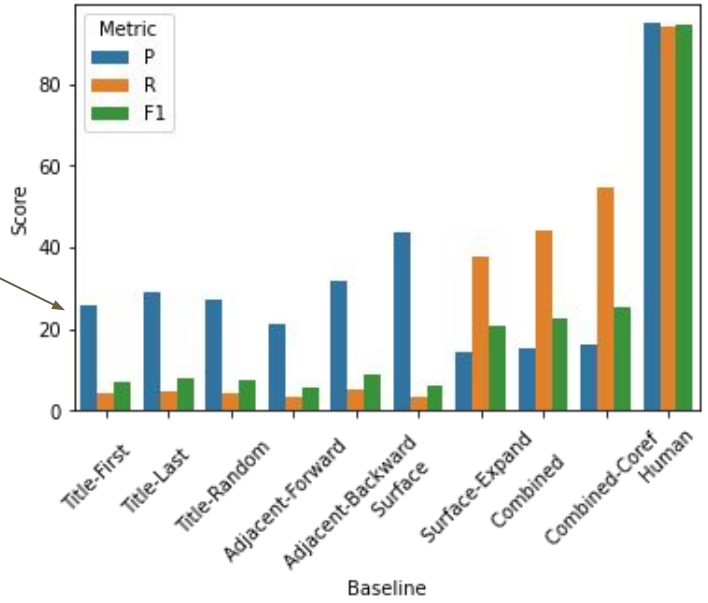
Crown Princess Mary of Denmark gives birth to male child

Her Royal Highness Crown Princess Mary of Denmark has given birth to a healthy baby boy at a Copenhagen hospital at approximately 1:57 am local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

TNE: Heuristic Baselines

- We live in a world of datasets biases
- Do we have any of these?

Link every NP to the **first** NP in the title



Title-First

<Denmark, **of**, Crown Princess Mary>
<birth, **of**, Crown Princess Mary>
<male child, **of** Crown Princess Mary>
…

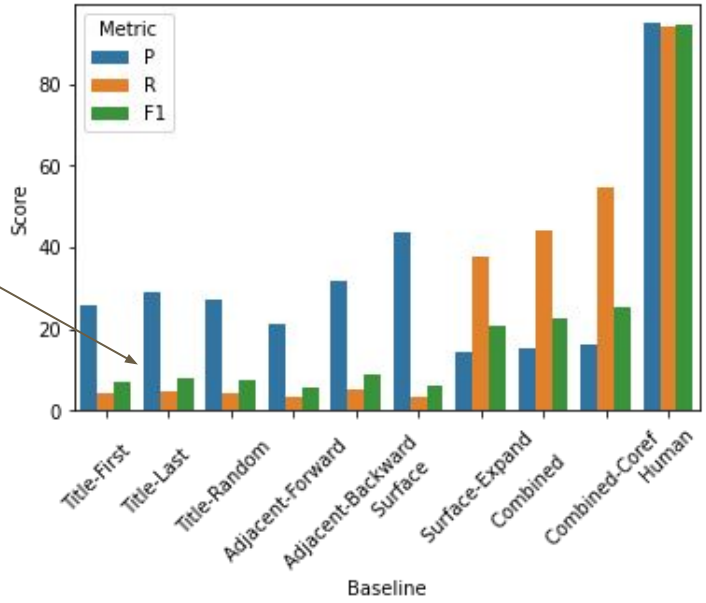
Crown Princess Mary of Denmark gives birth to male child

Her Royal Highness Crown Princess Mary of Denmark has given birth to a healthy baby boy at a Copenhagen hospital at approximately 1:57 am local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

TNE: Heuristic Baselines

- We live in a world of datasets biases
- Do we have any of these?

Link every NP to the **last** NP in the title



Title-Last

<Crown Princess Mary, **of**, male child>
<Denmark, **of**, male child>
<birth, **of**, male child>
…

Crown Princess Mary of Denmark gives birth to male child

Her Royal Highness Crown Princess Mary of Denmark has given birth to a healthy baby boy at a Copenhagen hospital at approximately 1:57 am local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

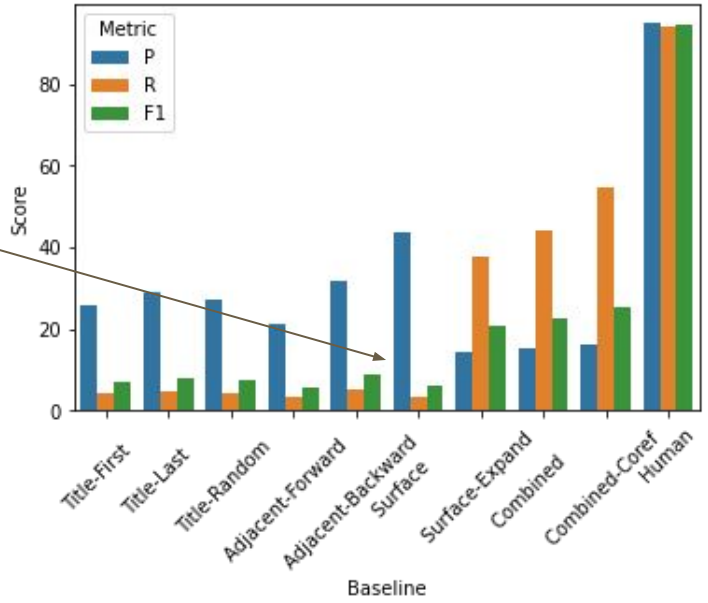
TNE: Heuristic Baselines

- We live in a world of datasets biases
- Do we have any of these?

Previous NP

<Denmark, **of**, Crown Princess Mary>
<birth, **of**, Denmark>
<male child, **of**,birth>
…

Link every NP to the **previous** NP in the text



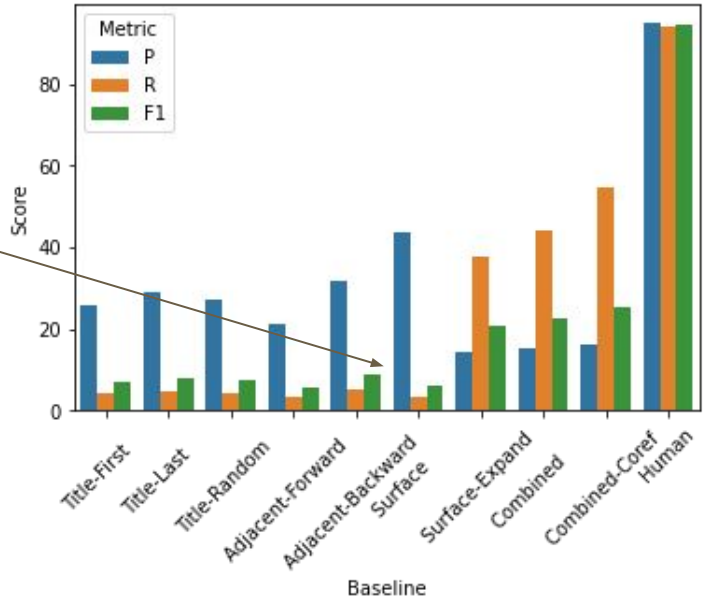
Crown Princess Mary of Denmark gives birth to male child

Her Royal Highness Crown Princess Mary of Denmark has given birth to a healthy baby boy at a Copenhagen hospital at approximately 1:57 am local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

TNE: Heuristic Baselines

- We live in a world of datasets biases
- Do we have any of these?

Link every NP to the **explicit**, surface form link



Explicit NP

<Crown Princess Mary, **of**, Denmark>
<birth, **to**, male child>
<waiting, **for**, the Royal Family>
…

Crown Princess Mary of Denmark gives birth to male child

Her Royal Highness Crown Princess Mary of Denmark has given birth to a healthy baby boy at a Copenhagen hospital at approximately 1:57 am local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

TNE: Heuristic Baselines

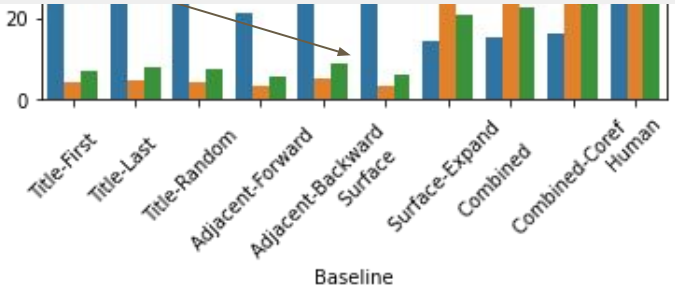
- We live in a world of datasets biases
- Do we have any of these?

Explicit NP

<Crown Princess Mary, **of**, Denmark>
<birth, **to**, male child>
<waiting, **for**, the Royal Family>
…

These heuristics can get you so far,

But it is very far from "human performance"



local time this morning, ending many months of waiting for the Royal Family, the Danish public and much of the world. The baby weighed in at 3.5 kilograms and 51 centimeters long.

TNE: Modeling + Results

- We proposed several baselines
- *More details in the paper*

Model	Precision	Recall	F1
Human*	94.8	94.0	94.4
Coupled-large	**65.8**	43.5	**52.4**

A big gap remains!

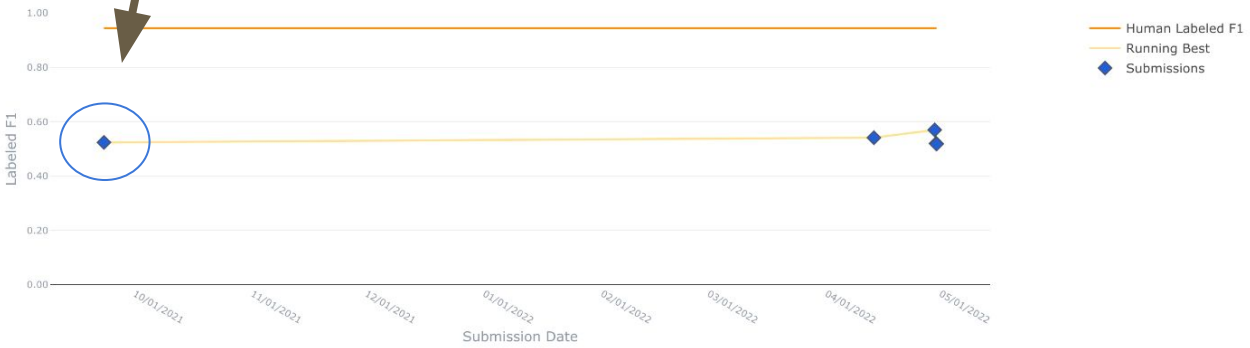
Leaderboards

Log In

Rank	Submission	Created	Labeled F1	Labeled Precision	Labeled Recall	Unlabeled F1	Unlabeled Precision	Unlabeled Recall
1	RoBERTa Large with closed-set... *Gal Fiebelman, Tom Turgeman, ...*	04/29/2022	**0.5697**	0.5300	**0.6158**	**0.6961**	0.6476	**0.7524**
2	Coupled Large with NP Attenti... *Adi Fine & Alon Mendelson fro...*	04/13/2022	0.5409	**0.6672**	0.4548	0.6598	**0.8139**	0.5548
3	Coupled SpanBERT-large *Yanai Elazar, Victoria Basmov...*	09/23/2021	0.5236	0.6582	0.4346	0.6401	0.8047	0.5314
4	End-to-end NP enrichment model *Yosi David from Tel Aviv Univ...*	04/29/2022	0.5199	0.6142	0.4507	0.6455	0.7626	0.5596
5	NP enrichment model *Yosi David from Tel Aviv Univ...*	04/29/2022	0.5196	0.6057	0.4550	0.6472	0.7545	0.5667
6	enhanced NP enrichment model *Yosi Cohen from Tel Aviv Univ...*	04/29/2022	0.5176	0.6153	0.4466	0.6405	0.7614	0.5527

This work

Labeled F1 Over Time



Leaderboards

Log In

Rank	Submission	Created	Labeled F1	Labeled Precision	Labeled Recall	Unlabeled F1	Unlabeled Precision	Unlabeled Recall
1	RoBERTa Large with closed-set... *Gal Fiebelman, Tom Turgeman, ...*	04/29/2022	**0.5697**	0.5300	**0.6158**	**0.6961**	0.6476	**0.7524**
2	Coupled Large with NP Attenti... *Adi Fine & Alon Mendelson fro...*	04/13/2022	0.5409	**0.6672**	0.4548	0.6598	**0.8139**	0.5548
3	Coupled SpanBERT-large *Yanai Elazar, Victoria Basmov...*	09/23/2021	0.[illegible]	0.6582	0.4346	0.6401	0.8047	0.5314
4	End-to-end NP enrichment model *[illegible]*	04/29/2022	[illegible]	0.6142	0.4507	0.6455	0.7626	0.5596
5	NP enrichment model *Yosi David from Tel Aviv Univ...*	04/29/2022	0.5196	0.6057	0.4550	0.6472	0.7545	0.5667
6	enhanced NP enrichment model *Yosi Cohen from Tel Aviv Univ...*	04/29/2022	0.5176	0.6153	0.4466	0.6405	0.7614	0.5527

Newer Models

This work

Labeled F1 Over Time

Leaderboards

Log In

Rank	Submission	Created	Labeled F1	Labeled Precision	Labeled Recall	Unlabeled F1	Unlabeled Precision	Unlabeled Recall
1	RoBERTa Large with closed-set... *Gal Fiebelman, Tom Turgeman, ...*	04/29/2022	**0.5697**	0.5300	**0.6158**	**0.6961**	0.6476	**0.7524**
2	Coupled Large with NP Attenti... *Adi Fine & Alon Mendelson fro...*	04/13/2022	0.5409	**0.6672**	0.4548	0.6598	**0.8139**	0.5548
3	Coupled SpanBERT-large *Yanai Elazar, Victoria Basmov...*	09/23/2021	0.5[illegible]	0.6582	0.4346	0.6401	0.8047	0.5314
4	End-to-end NP enrichment model *[illegible]*	04/29/2022	[illegible]	0.6142	0.4507	0.6455	0.7626	0.5596
5	NP enrichment model *Yosi David from Tel Aviv Univ...*	04/29/2022	0.5196	0.6057	0.4550	0.6472	0.7545	0.5667
6	enhanced NP enrichment model *Yosi Cohen from Tel Aviv Univ...*	04/29/2022	0.5176	0.6153	0.4466	0.6405	0.7614	0.5527

Newer Models

This work

Labeled F1 Over Time

Still a Hard Task!

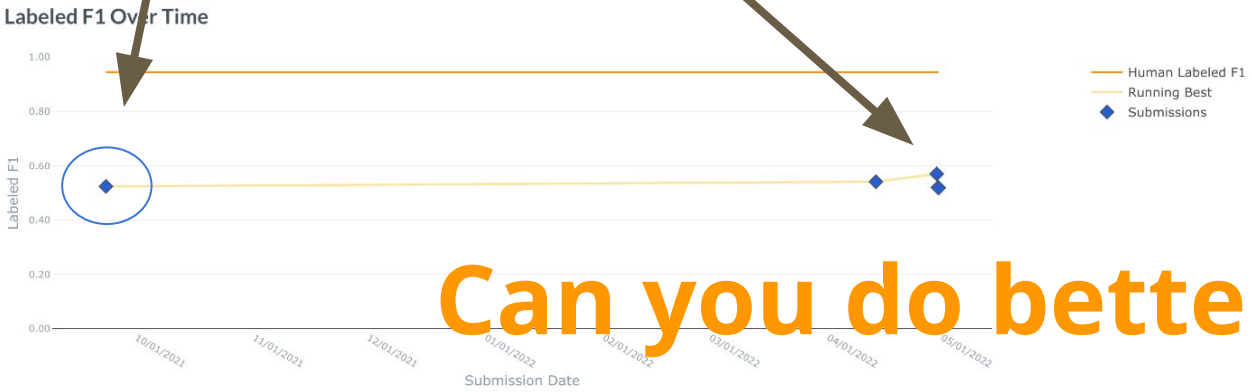
Leaderboards

Log In

Rank	Submission	Created	Labeled F1	Labeled Precision	Labeled Recall	Unlabeled F1	Unlabeled Precision	Unlabeled Recall
1	RoBERTa Large with closed-set... *Gal Fiebelman, Tom Turgeman, ...*	04/29/2022	**0.5697**	0.5300	**0.6158**	**0.6961**	0.6476	**0.7524**
2	Coupled Large with NP Attenti... *Adi Fine & Alon Mendelson fro...*	04/13/2022	0.5409	**0.6672**	0.4548	0.6598	**0.8139**	0.5548
3	Coupled SpanBERT-large *Yanai Elazar, Victoria Basmov...*	09/23/2021	0.[illegible]	0.6582	0.4346	0.6401	0.8047	0.5314
4	End-to-end NP enrichment model	04/29/2022	[illegible]	0.6142	0.4507	0.6455	0.7626	0.5596
5	NP enrichment model *Yosi David from Tel Aviv Univ...*	04/29/2022	0.5196	0.6057	0.4550	0.6472	0.7545	0.5667
6	enhanced NP enrichment model *Yosi Cohen from Tel Aviv Univ...*	04/29/2022	0.5176	0.6153	0.4466	0.6405	0.7614	0.5527

Newer Models

This work



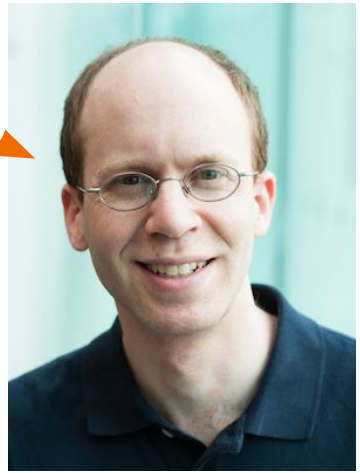
Still a Hard Task!

Can you do better?

Story #2

- In the paper, we also cover other stories:
  - Recovering implicit information
  - Crowdsourcing a large, high-quality dataset
  - Extensive comparison to other linguistic phenomena

Including an answer for



"**The Use of Prepositions as Semantic Labels** While the relations we identify between NPs can be expressed using prepositions, one could argue that using prepositions as semantic labels is not ideal, due to their inherent ambiguity (Schneider et al., 2015, 2016, 2018; Gessler et al., 2021): indeed a preposition such as *for* has multiple senses, and can indicate a large set of semantic relations ranging from BENEFICIARY to DURATION."

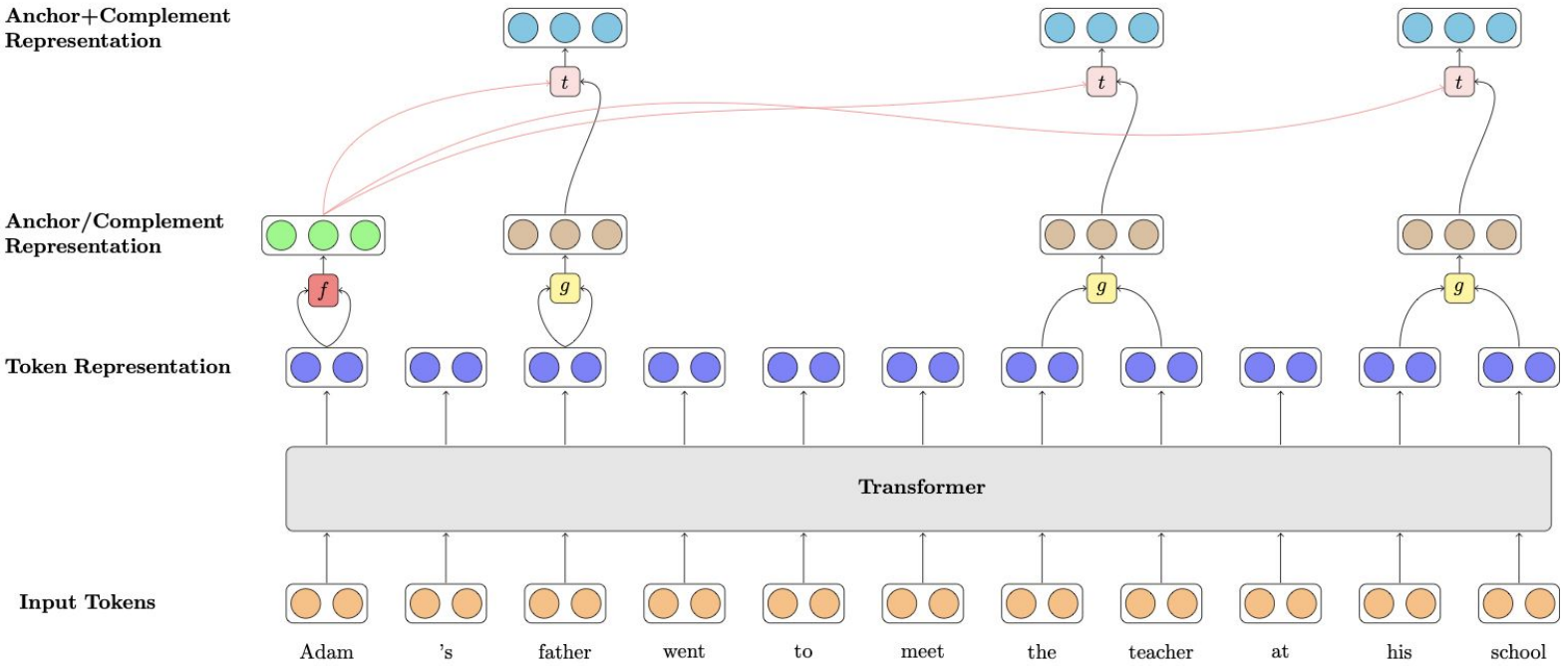
Story #2

- In the paper, we also cover other stories:
  - Recovering implicit information
  - Crowdsourcing a large, high-quality dataset
  - Extensive comparison to other linguistic phenomena
- Check it out!

Summary

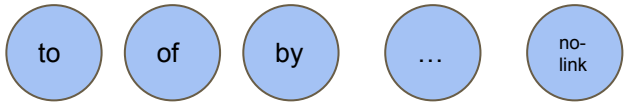
- TNE: a new benchmark for **Reading Comprehension**
- ...and a useful task, with many applications: Relation Extraction, Question Answering, etc.
- A large, high-quality dataset
- A challenging task

TNE: Modeling

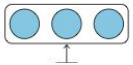


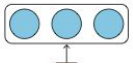
TNE: Modeling

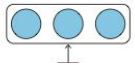
- 2.5 Variants:
  - **Coupled**: A head that predicts 1 of the 23 prepositions or *no-link*



Anchor+Complement Representation

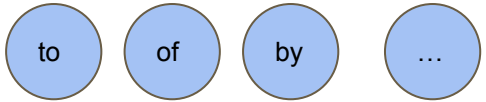


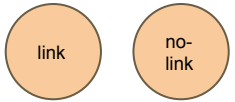




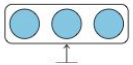
TNE: Modeling

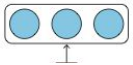
- 2.5 Variants:
  - **Coupled**: A head that predicts 1 of the 23 prepositions or *no-link*
  - **Decoupled**: 1 head that predicts if a link exists **and...**
    1 head that predicts the preposition (if it exists)

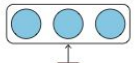




Anchor+Complement
Representation



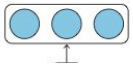


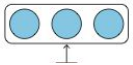


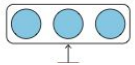
TNE: Modeling

- 2.5 Variants:
  - **Coupled**: A head that predicts 1 of the 23 prepositions or *no-link*
  - **Decoupled**: 1 head that predicts if a link exists **and...**
    1 head that predicts the preposition (if it exists)
  - **Static**: both versions, but freezing the encoder model (similar to probing)

Anchor+Complement Representation







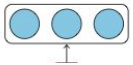
TNE: Results

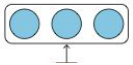
- 2.5 Variants:
  - **Coupled**: A head
  - **Decoupled**: 1 head
    1 head
  - **Static**: both versions, b

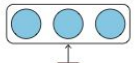
	Model	Precision	Recall	F1
	Human*	94.8	94.0	94.4
Pretrained	Decoupled-static	10.1	58.8	17.2
	Decoupled-frozen-base	9.6	55.5	16.3
	Decoupled-frozen-large	9.7	56.2	16.5
	Decoupled-base	11.8	68.5	20.1
	Decoupled-large	12.0	**69.9**	20.5
	Coupled-static	59.6	14.4	23.2
	Coupled-frozen-base	60.1	8.6	15.1
	Coupled-frozen-large	58.4	11.5	19.2
	Coupled-base	60.4	41.5	49.2
	Coupled-large	**65.8**	43.5	**52.4**

)

Anchor+Complement Representation







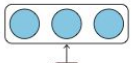
TNE: Results

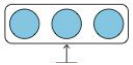
- 2.5 Variants:
  - **Coupled**: A head
  - **Decoupled**: 1 head
    1 head
  - **Static**: both versions, b

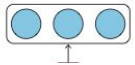
	Model	Precision	Recall	F1
	Human*	94.8	94.0	94.4
Pretrained	Decoupled-static	10.1	58.8	17.2
	Decoupled-frozen-base	9.6	55.5	16.3
	Decoupled-frozen-large	9.7	56.2	16.5
	Decoupled-base	11.8	68.5	20.1
	Decoupled-large	12.0	**69.9**	20.5
	Coupled-static	59.6	14.4	23.2
	Coupled-frozen-base	60.1	8.6	15.1
	Coupled-frozen-large	58.4	11.5	19.2
	Coupled-base	60.4	41.5	49.2
	Coupled-large	**65.8**	43.5	**52.4**

A big gap remains!

Anchor+Complement Representation

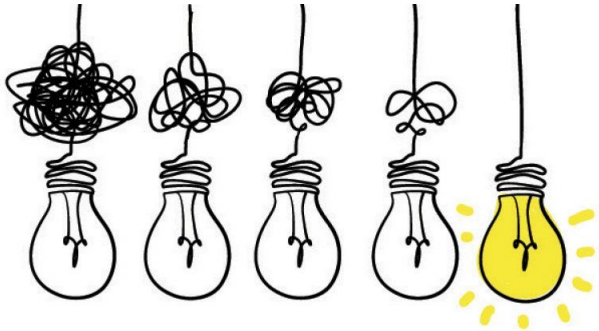






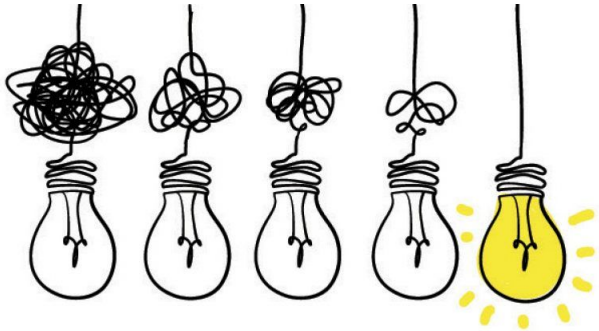
Bonus: TNE, A 3 Year Long Project

- Simplicity



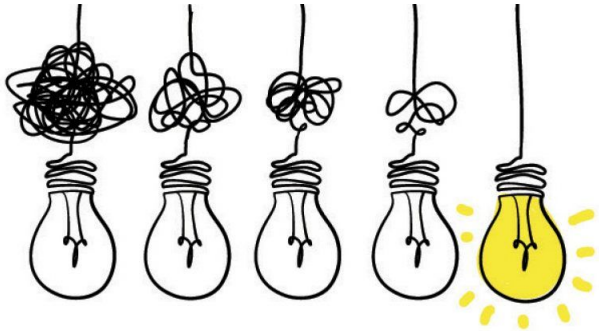
Bonus: TNE, A 3 Year Long Project

- Simplicity
- Quality



Bonus: TNE, A 3 Year Long Project

- Simplicity
- Quality
- Availability



End of Part I

Questions?

```
from datasets import load_dataset

dataset = load_dataset("tne")
```

SCAN ME

yanaiela.github.io/TNE/

github.com/yanaiela/TNE

leaderboard.allenai.org/tne/

But What About ChatGPT???

Task Definition: The Text-based NP Enrichment task (TNE) is deceptively simple: for each ordered pair (n1, n2) of non-pronominal base-NP spans in an input text, determine if there exists a preposition mediated relation between n1 and n2, and if there is one, determine the preposition that best describes their relation.4 The output is a list of tuples of the form (ni, prep, nj), where ni is called the anchor and nj is called the complement of the relation. Figure 2 shows an example of text where each NP n1 is annotated with its (prep, n2) NP enrichments.

Task Definition: The Text-based NP Enrichment task (TNE) is deceptively simple: for each ordered pair (n1, n2) of non-pronominal base-NP spans in an input text, determine if there exists a preposition mediated relation between n1 and n2, and if there is one, determine the preposition that best describes their relation.4 The output is a list of tuples of the form (ni, prep, nj), where ni is called the anchor and nj is called the complement of the relation. Figure 2 shows an example of text where each NP n1 is annotated with its (prep, n2) NP enrichments.

The Text-based NP Enrichment (TNE) task involves analyzing a given text to identify preposition-mediated relations between non-pronominal base-NP spans. The task requires determining if th preposition that best describes t here each tuple contains an anch P is the first NP in the relation, the complement NP is the second NP, and the preposition mediates the relation between them. The TNE task is an important step towards deeper natural language understanding and can help improve various natural language processing applications such as information extraction, question-answering, and text summarization.

Who asked you??

"UN agrees on reform document
September 17, 2005
A deal to reform the United Nations has been signed at the end of a summit marking the organisation's sixtieth anniversary.
The summit was the largest gathering of heads of state ever, with over 150 in attendance.
During its three days, Syria was the only one of the 191 represented nations not to address the General Assembly.
The Secretary General, Kofi Annan, had put forward a wide-ranging set of reform proposals encompassing poverty, war, genocide, and reform of the organisation itself."

Generate the TNE triplets for "The Secretary General"

"UN agrees on reform document
September 17, 2005
A deal to reform the United Nations has been signed at the end of a summit marking the organisation's sixtieth anniversary.
The summit was the largest gathering of heads of state ever, with over 150 in attendance.
During its three days, Syria was the only one of the 191 represented nations not to address the General Assembly.
The Secretary General, Kofi Annan, had put forward a wide-ranging set of reform proposals encompassing poverty, war, genocide, and reform of the organisation itself."

Generate the TNE triplets for "The Secretary General"

There are no preposition-mediated relations mentioned in the given text for the NP span "The Secretary General", so it is not possible to generate TNE triplets for it.

"UN agrees on reform document
September 17, 2005
A deal to reform the United Nations has been signed at the end of a summit marking the organisation's sixtieth anniversary.
The summit was the largest gathering of heads of state ever, with over 150 in attendance.
During its three days, Syria was the only one of the 191 represented nations not to address the General Assembly.
The Secretary General, Kofi Annan, had put forward a wide-ranging set of reform proposals encompassing poverty, war, genocide, and reform of the organisation itself."

Generate the TNE triplets for "The Secretary General"

There are no preposition-mediated relations mentioned in the given text for the NP span "The Secretary General", so it is not possible to generate TNE triplets for it.

How about:
"The Secretary General of the UN"?

Wait What?

"UN agrees on reform document
September 17, 2005
A deal to reform the United Nations has been signed at the end of a summit marking the organisation's sixtieth anniversary.
The summit was the largest gathering of heads of state ever, with over 150 in attendance.
During its three days, Syria was the only one of the 191 represented nations not to address the General Assembly.
The Secretary General, Kofi Annan, had put forward a wide-ranging set of reform proposals encompassing poverty, war, genocide, and reform of the organisation itself."

Generate the TNE triplets for "The Secretary General"

There are no preposition-mediated relations mentioned in the given text for the NP span "The Secretary General", so it is not possible to generate TNE triplets for it.

How about:
"The Secretary General of the UN"?

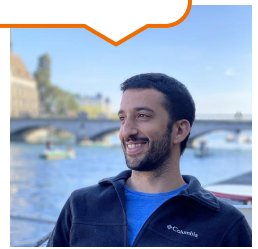
For the NP span "The Secretary General of the UN", there is one preposition-mediated relation mentioned in the given text:

- Anchor NP: The Secretary General of the UN
- Preposition: had put forward
- Complement NP: a wide-ranging set of reform proposals encompassing poverty, war, genocide, and reform of the organisation itself

Therefore, the TNE triplet for "The Secretary General of the UN" in the given text is:

("The Secretary General of the UN", "had put forward", "a wide-ranging set of reform proposals encompassing poverty, war, genocide, and reform of the organisation itself")

Wait What?



What About Other Missing Elements?

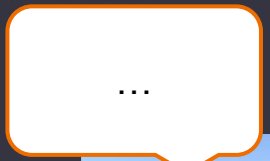
- TNE
- Fused-Heads
- Complement Coercion

Numeric fused heads are numbers that were fused with their dependent modifiers. For instance, in the sentence "I'm 42, Cercie", the number 42 was fused with the implicit age indicator "years-old".

What is the missing head in "It's worth about two million"?

Numeric fused heads are numbers that were fused with their dependent modifiers. For instance, in the sentence "I'm 42, Cercie", the number 42 was fused with the implicit age indicator "years-old".

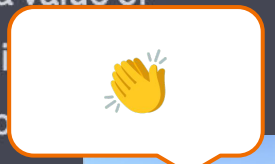
What is the missing head in "It's worth about two million"?

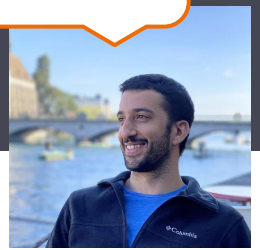


In the sentence "It's worth about two million," the missing head is the noun or objec
phrase "two million" modifies. Without additional context, it is unclear what exactly is worth two million. It could be a house, a car, a business, or any other valuable asset.

What does two million refers to in "It's worth about two million"?

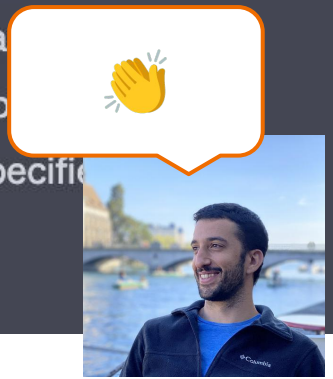
Without additional context, it is unclear what exactly the phrase "two million" refers to in the sentence "It's worth about two million." The sentence implies that something has a value of approximately two million, but it does not specify what that something is. Dependi
context, "two million" could refer to dollars, euros, pounds, or any other currency, o
refer to some other type of value such as shares, units, or items.





What does "one" refers to in: "Jenny: Head in the clouds, that one. I don't know why you're so sweet on him"?

In the sentence "Jenny: Head in the clouds, that one. I don't know why you're so sweet on him," the word "one" is used as a pronoun to refer to a person, specifically the person that Jenny is talking about. The sentence implies that the person being referred to ha
tendency to daydream or not pay attention to their surroundings, as evidenced b
phrase "head in the clouds." The identity of the person being referred to is not specifi
this sentence.



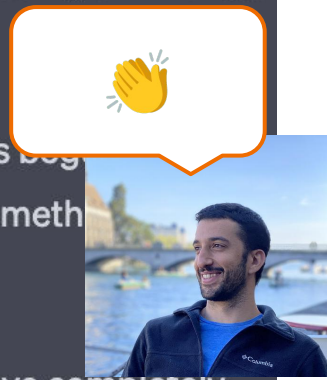
What's the relationship between the following:
I started a book
I started eating a book

The two sentences are related in that they both involve the verb "started" and the noun "book", but the meaning of each sentence is very different.

The first sentence, "I started a book", means that the speaker has begun reading or writing a book. It implies that the speaker is engaged in a literary activity.

The second sentence, "I started eating a book", means that the speaker has beg consume a book, as if it were food. This sentence is nonsensical and not someth would typically happen in reality.

Therefore, while the two sentences share some common elements, they have completely different meanings and contexts.



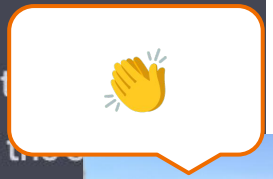
I started a book
I started reading a book

The relationship between the two sentences is that the second sentence is a specific type of the first sentence, and the relationship can be labeled as "entailment".

Specifically, the first sentence, "I started a book", is a general statement that the speaker has started some activity related to a book, but it doesn't specify what that activity is.

The second sentence, "I started reading a book", is a more specific statement t
more information about the activity the speaker has started - specifically, that t
has started reading the book.

Therefore, the second sentence "I started reading a book" entails the first sentence
a book", because if the second sentence is true, then it logically follows that the first
sentence is also true.



How about
I started a book
I started writing a book

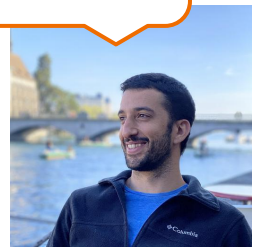
The relationship between the two sentences is that they are both examples of starting an activity related to a book, but they are not necessarily the same activity. The relationship can be labeled as "neutral" or "contradiction", depending on the context.

If the context implies that the speaker is a writer, then the relationship between the two sentences can be labeled as "entailment". Specifically, the first sentence "I started a book" implies that the speaker has begun work on a book in some way, and the second sentence "I started writing a book" provides more specific information about the type of work that the speaker has started on the book.

However, if the context implies that the speaker is a reader, then the relationship between the two sentences can be labeled as "contradiction". Specifically, the first sentence "I started a book" implies that the speaker has begun reading a book, while the second sentence "I started writing a book" implies that the speaker has begun authoring a book. These two activities are not compatible, so the two sentences would be in contradiction with each other.

Therefore, the relationship between the two sentences can be either "entailment" or "contradiction", depending on the context. In a neutral context where no additional information is given, the relationship between the two sentences would be considered "neutral", as there is not enough information to determine if they are logically related.

...



ChatGPT and Missing Elements

- The understanding of such constructions is still not solved
  - Although it does work surprisingly (?) well out of the box
- But let’s look back at why these questions are interesting

ChatGPT and Missing Elements

- A more traditional approach to make things work
- Part of the “NLP Pipeline”
  - Part-of-speech
  - Coreference resolution
  - Syntactic parsing
  - Missing elements?
  - ...
- Are these tasks still relevant?

ChatGPT and Traditional NLP Tasks

- Are these tasks still relevant (as of 2023)?

Yes, if:

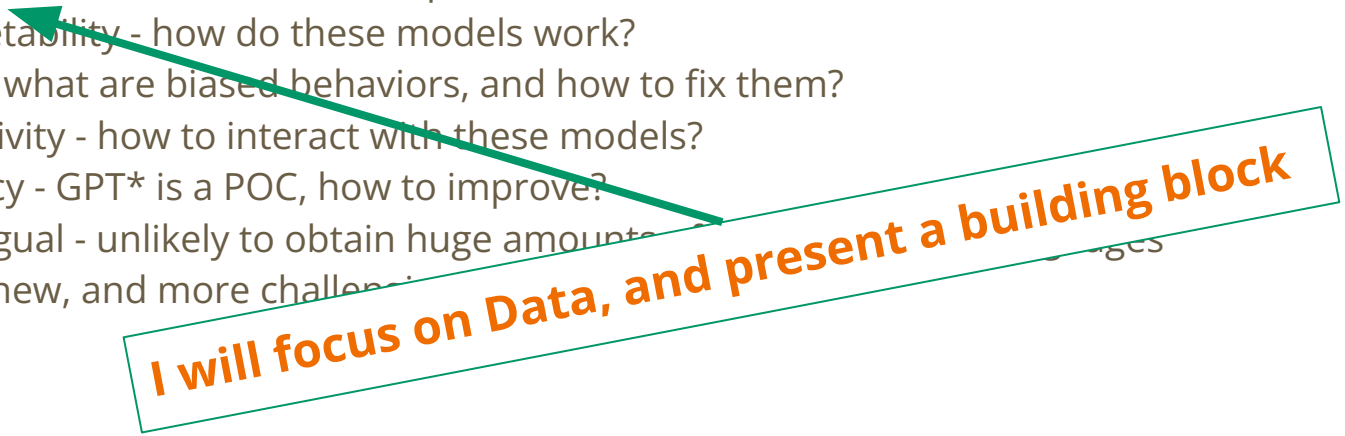
- You are interested in language
- You work with low resources languages
- You want to extract information from the entire web

No, if:

- You want to make things work (and you can afford the large-scale models)
- You are interested in AGI

ChatGPT and Traditional NLP Tasks

- “Somewhat useless to model such tasks in the age of GPT-3”
- So what’s left to research?
  - Datasets - what in the data is responsible for LLM’s behavior?
  - Interpretability - how do these models work?
  - Biases - what are biased behaviors, and how to fix them?
  - Interactivity - how to interact with these models?
  - Efficiency - GPT* is a POC, how to improve?
  - Multilingual - unlikely to obtain huge amounts
  - Tasks - new, and more challen
  - …



WIMBD: What's In My Big Data?

Scale of Data

	A Neural Probabilistic Language Model	*NLP (Almost) From Scratch*	*Word2Vec*	*ELMo*	*BERT*	*GPT2*	*GPT3*	*BLOOM*	*LLaMA*
	Bengio et al 2003	Collobert et al 2011	Mikolov et al 2013	Peters et al 2018	Devlin et al 2018	Radford et al 2019	Brown et al 2020	Big Science 2022	Touvron et al 2023
PARAMS	1.2M	5M	300M*	93M	330M	1.5B	175B	175B	65B
TOKENS	14M	631M	1.6B	1.8B	3.3B	~14B	~400B	1.6T*	1.4T

Thanks Luca Soldaini for the slide!

Scale of Data

- So how to deal with this much data?
- Remember BigData from 10 years ago?
  - Let’s reuse these ideas

WIMBD: High Level Capabilities

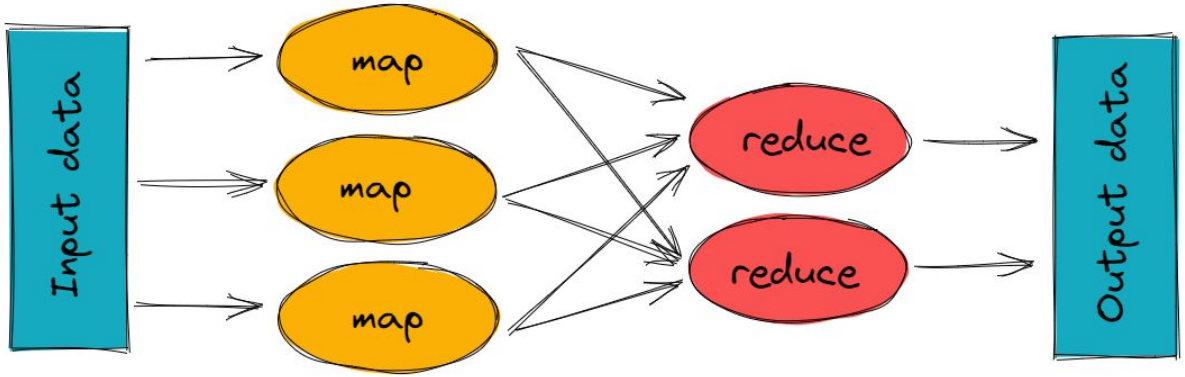
- *A toolkit to study textual data*
- **Counts**
  - Exact Counts
  - Compressed Counts
  - Estimated Counts
  - Sampled Counts
- **Queries**
  - Retrieval

WIMBD: Counts

- Exact Counts
  - Map-Reduce

We can get:

- Domain distribution
- Length distribution
- ...



WIMBD: Counts

- Compressed Counts
  - Dictionary (a big one)

We can get:

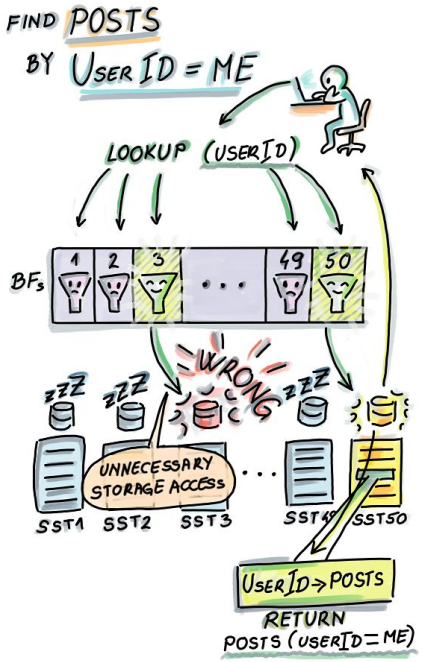
- Duplicates
- Near duplicates

WIMBD: Counts

- Estimated Counts
  - Bloom Filters

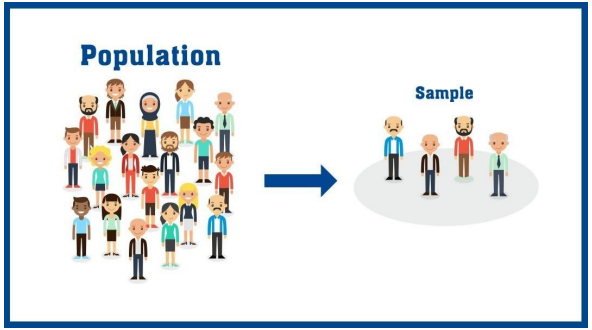
We can get:

- N-gram counts



WIMBD: Counts

- Sampled Counts
  - Random access



We can get:

- Time-consuming operations

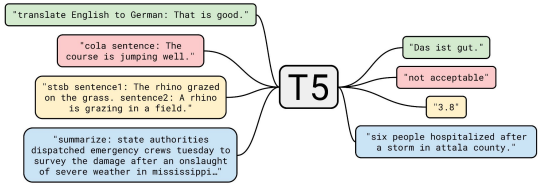
WIMBD: Retrieval

- Retrieving matching documents
  - Elastic-Search

We can get:

- Dataset contamination
- Your personal data
- ...

WIMBD: Analyzing Datasets



- C4
- The Pile
- Oscar
- S2ORC
- OpenWebText
- LAION
- The Stack

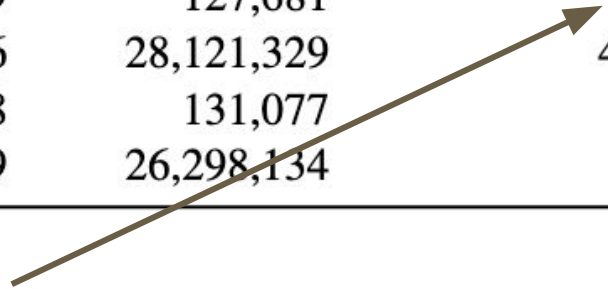
GPT-j / GPT-neo / phythia

GPT-2

Code but - “... In python files, [English] makes up to ~96% of the dataset

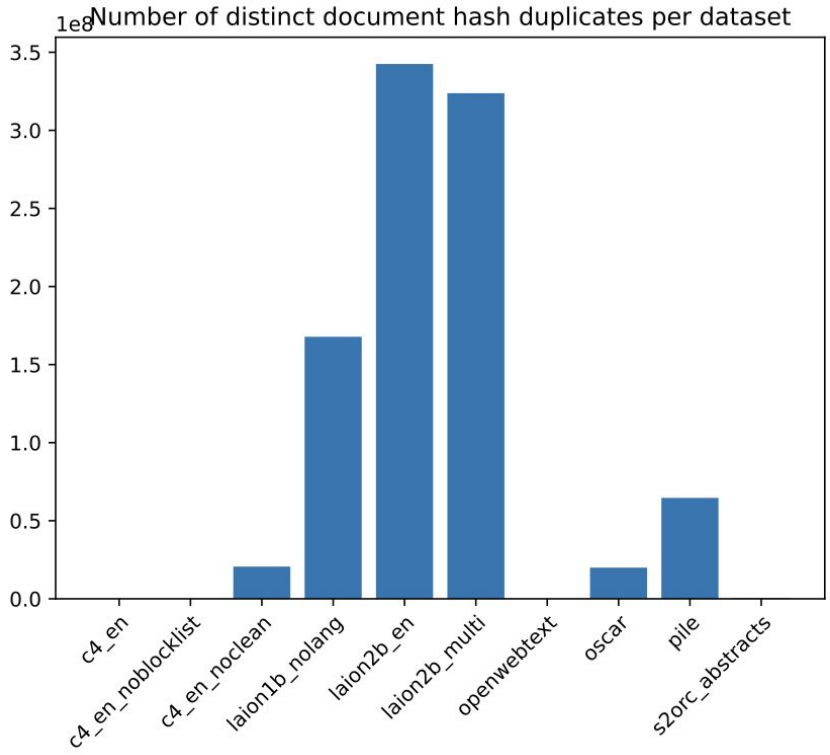
WIMBD: A Few Interesting Results: Basic Stats

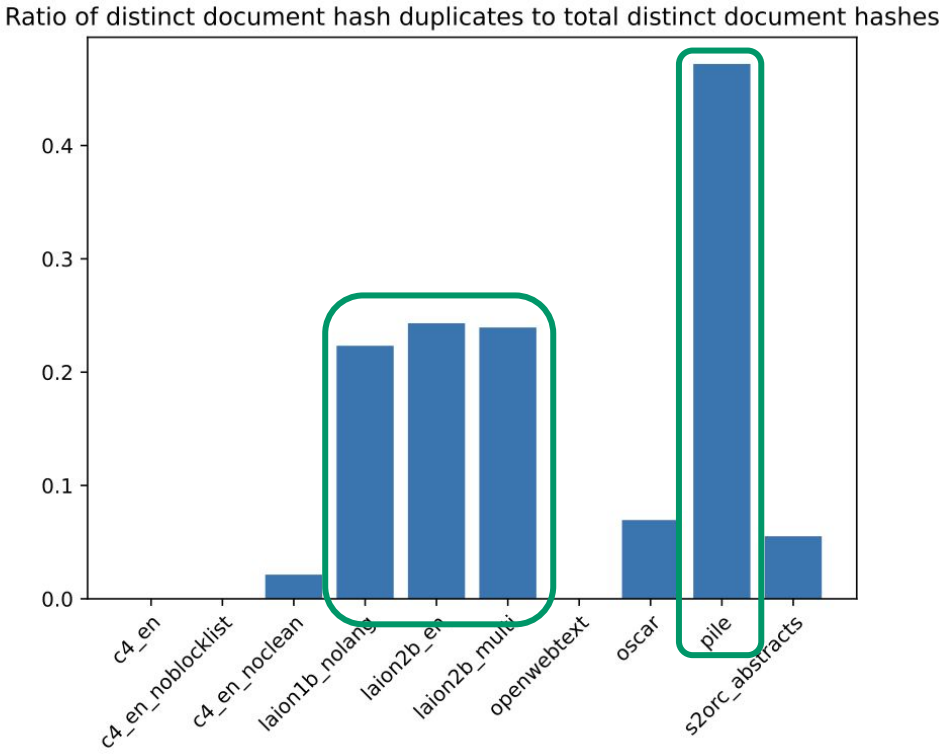
Dataset	Size (Gb)	# Documents	# Tokens	max(# Tokens)	min(# Tokens)
c4	838.7	364,868,892	153,607,833,664	101,898	12
openwebtext	40.6	8,013,769	7,775,216,373	95,139	137
oscar	3,205.7	431,992,659	476,431,685,530	1,048,409	12
s2orc	18.6	31,050,107	2,882,388,199	127,681	0
pile	1,369.0	210,607,728	285,794,281,816	28,121,329	48
laion2B	454.6	2,322,161,808	29,672,214,438	131,077	1
stack	7,882.2	545,283,351	1,527,098,679,269	26,298,134	1



Always check your data!

WIMBD: A Few Interesting Results: Duplicates





WIMBD: A Few Interesting Results: Duplicates

laion2b en		laion2b multi		laion1b nolang		s2orc abstracts	
text	count	text	count	text	count	text	count
Front Cover	1003863	Bathroom	490003	CAPTCHA	3377045		20948552
Wall View 002	681753	Kitchen	441649	CAPTCHA Image	1853156	Abstract	35716
Market position of the selected technologies	414986	Framed Print	438812	Google Play	1726671	ABSTRACT	25832
Pointwise: Reliable CFD meshing	319524	Tri-blend T-Shirt	362399	European Commission logo	688732	Background	18055
Go to European Commission website	314423	Bedroom	362091	Screenshot Image	531195	Abstract\nBackground	4486
The article as it originally appeared.	311739	Classic T-Shirt	346799	Real Estate Photo	315161	The	2979
Patent Drawing	294928	avatar	344368	Property Photo	312724	Abstract-	2395
Cyprus property for sale	293536	iPhone Case/Skin	299168	Women's T-Shirt	290660	Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. [...]	2295
Throw Pillow	278756	CAPTCHA изображение	274498	Imagen CAPTCHA	277353	Introduction	1918
Wall View 003	235226	Lightweight Sweatshirt	223662	Floorplan	274573	Abstract\nIntroduction	1736

WIMBD: A Few Interesting Results: Contamination

Dataset	Split	Field	C4	openwebtext	Oscar	Pile
wnli	validation	sentence1	26.76	9.86	30.99	32.39
		sentence2	16.9	2.82	15.49	14.08
	test	sentence1	4.79	0.0	0.0	4.79
		sentence2	2.05	0.68	2.74	2.74
stsb	validation	sentence1	15.87	1.8	13.87	36.0
		sentence2	16.2	2.47	13.2	37.33
	test	sentence1	25.02	9.06	21.32	35.75
		sentence2	25.82	8.41	20.38	36.84
sst2	validation	sentence	14.33	1.61	7.22	17.43
	test	sentence	14.61	2.36	7.96	20.26
rte	validation	sentence1	8.66	1.44	8.3	9.03
		sentence2	15.88	2.89	14.44	18.41
	test	sentence1	9.27	1.0	8.07	5.2
		sentence2	13.17	4.37	11.8	10.13
qqp	validation	question1	17.46	2.31	13.77	7.64
		question2	18.05	2.67	14.42	8.37
qnli	validation	question	2.31	0.51	1.89	4.03
		sentence	83.86	43.97	69.27	76.92
	test	question	2.45	0.37	1.78	4.14
		sentence	84.15	42.58	69.6	76.5

WIMBD: A Few Interesting Results: Toxic Language

Corpus	taxonomy	classifier
C4	0.01	0.17
Openwebtext	13.8	0.8
Oscar	8.98	0.57
Laion2B-en	0.89	1.01
Pile	7.67	0.7
S2orc-abstracts	2.58	0.06
stack	1.85	0.07

Table 6: Toxic language percentages based on a taxonomy and a classifier over entire documents in the corpora we consider

Zhou et. al 2021

WIMBD: Summary

- To understand models - *What*, *How*, *Why*:
- **We need to understand the data**
- We can’t claim:
  - ***“Emergent abilities”*** without understanding the data
  - ***Generalization*** without de-contaminating the data
  - ***Bias Amplification*** without measuring biases in the data
  - ***Memorization*** without testing for presence in the data
  - ...



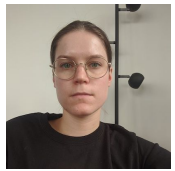

TBD


WIMBD: Summary

- Provide a set of tools that allows studying large corpora
- Analyze seven corpora, of different nature, some used for training LMs
- Can be seen as filtering advice, as well as post-hoc studies





End of Part 2

Thank you
Questions?

@yanaiela

yanaiela.github.io