



# LongtoNotes: OntoNotes with Longer Coreference Chain

Kumar Shridhar, Nicholas Monath, Raghuveer Thirukovalluru, Alessandro Stolfo,  
Manzil Zaheer, Andrew McCallum, Mrinmaya Sachan

ETH Zurich, Umass Amherst, Duke University, Google Deepmind

# Introduction

- **OntoNotes** has served as one of the most **important benchmark** for coreference resolution.

# Introduction

- **OntoNotes** has served as one of the most **important benchmark** for **coreference resolution**.
- **# of documents: 3493**
- **Total size (tokens): 1.6M**
- **Tokens per document: 466**

# However

- For the ease of annotation, **several long documents in Ontonotes were split** into smaller parts.

# However

- For the ease of annotation, **several long documents in Ontonotes were split** into smaller parts.

The flight attendants have until 6:00 today to ratify labor concessions. The pilots' union and ground crew did so yesterday.

...

The airlines provided report on these pilots and decided to file a case against them.

# However

- For the ease of annotation, **several long documents in Ontonotes were split** into smaller parts.
- We **cannot precisely test the performance** of models on the entire OntoNotes corpus.

# LongtoNotes

- As the name suggests, **we propose a longer version of the OntoNotes corpus.**

# LongtoNotes

- As the name suggests, **we propose a longer version of the OntoNotes corpus.**
- Upto **8X longer documents**



# LongtoNotes

- As the name suggests, **we propose a longer version of the OntoNotes corpus.**
- Upto **8X longer documents**
  - On an average **40% longer documents**

# LongtoNotes

- As the name suggests, **we propose a longer version of the OntoNotes corpus.**
- Upto **8X longer documents**
  - On an **average 40% longer documents**
- On an average **25% longer coreference chains**

# LongtoNotes

- As the name suggests, **we propose a longer version of the OntoNotes corpus.**
- Upto **8X longer documents**
  - On an **average 40% longer documents**
- On an average **25% longer coreference chains**
- **30% more mentions** per coreference chain

# Annotation Process

- We **merge the annotations** from documents **that were split** into multiple parts in the original Ontonotes annotation process, **into one single document.**

# Annotation Process

- We **merge the annotations** from documents **that were split** into multiple parts in the original Ontonotes annotation process, **into one single document**.
- This was done by our **annotation team**, which was carefully trained to follow the annotation guidelines (based on the original OntoNotes corpus).
  - Total annotation time: **400 hours**

# Why not do the merging automatically

- We tried **automatic merging** of clusters using:
  - **Rule based matching** : Stanford coref system
  - **Greedy matching**
  - **Model based**

# Why not do the merging automatically

- We tried **automatic merging** of clusters using:
  - **Rule based matching** : Stanford coref system
  - **Greedy matching**
  - **Model based**
- **Only 60-70% overlap with the annotators**
  - Not good enough
    - Especially for longer coref chains

# Annotation Tool

- **Changed merging two chains into a question answering task**
  - Annotators needs to merge two chains into one or assign a 'None of the previous' label (new chain option).



# Annotation Tool

1 Terracotta Warriors invade Taipei -LRB- Tsai Wen - ting / tr . by David Mayer -RRB-

4 The terracotta warriors from the tomb of Qin Shi Huang , which is included on the list of World Heritage sites of the UNESCO , have come ashore in 3 Taiwan !

1 The 17 warriors now on display at 4 the National Museum of History represent the largest contingent from the ancient tomb ever to travel on loan outside of mainland China . `` Reporting to the general , sir !

5 A mob of thousands has surrounded the building ! `` 6 A mounted warrior delivers the dire news to 7 6 his commander , whose high rank can be seen in 7 his tasseled helmet and coat of mail.

5 The `` mob , `` of course , is thirsty not for blood . 5 They just want a glimpse of 1 the warriors . Although 1 the warriors may not be aware of 8 it , 1 they 8 are the star attraction at

9 a blockbuster exhibit showing now at 4 the National Museum of History in 2 Taipei . Since opening on December 15 , 9 the exhibit has touched off a wave of enthusiasm in 3 Taiwan for

1 the terracotta warriors , and lines for admission have often snaked along the street for hundreds of meters . Over 20,000 advance tickets have been sold , and group visits are already booked solid to the end of January .

10 Many schools in central and southern Taiwan are renting buses by the dozen to ferry 10 their students up to 2 Taipei , and lots of foreign tourists are also coming by for a look .

The highlighted concept below refers to which concept in the above paragraphs?

Eight wonder Shaanxi Province was hit by a long drought in the summer of 1974. The farmer Yang Zhifa and a neighbor were digging a well when a shovel clinked on something hard. Yang exclaimed, `` Whoa ! It's them ! `` Locals had talked for generations about the buried clay figures that they would sometimes run across while digging in the fields. Everyone knew they were out there, but no one knew quite what they were. People figured they must have been connected with ancestral spirits of some sort. Anxious not to run afoul of the netherworld, the farmers would always hurriedly throw dirt back over the spot and find a better place to dig. This time, however, Yang Zhifa went a step further. He reported the incident to his superiors at the local commune. It was n't long before archeologists were swarming over the site, unearthing **an army of terracotta warriors that had lain buried for over 2,000 years**. **The stunning discovery** was dubbed the eighth wonder of the world. **The terracotta army** gives a graphic illustration of the grand scale of Qin Shi Huang's ambitions. **It** also shows how technologically advanced China already was when Qin Shi Huang was buried some 2,200 years ago. In order to create the feel of an actual descent into the emperor's tomb, the museum has arranged the exhibit so that visitors enter on the second floor and descend from there into a tomb - like atmosphere on the first floor, where the dimly lit space affords a glimpse of what was on the emperor's mind during the years he was busily uniting China for the very first time in history.

Answers

None of the Below

1 Terracotta Warriors

1 The 17 warriors now on display at the National Museum of History

5 The `` mob 5 A mob of thousands

8 it 8 are

10 Many schools in central and southern Taiwan 10 the

SUBMIT ANSWER

# Annotation Tool

- **Changed merging two chains into a question answering task**
  - Annotators needs to merge two chains into one or assign a 'None of the previous' label (new chain option).
- **Color coordination with numbers for better labelling**
  - Annotators **can change answers anytime**
  - Can pause and continue
  - Summary page to check for one final time

# Summary Page

## Summary of Your Answers

### Highlighted Mention

7 the Eighth Route Army ; 7 the Eighth Route Army ; 7 Eighth Route Army ;  
 7 the Eighth Route Army ; 7 the Eighth Route Army ; 7 Eighth Route Army ;  
 7 the Eighth Route Army led by the Chinese Communist Party ; 7 the Eighth Route Army ; 7 its

3 the guerrillas ; 3 The Eighth Route Army guerrillas

### Answer

7 the Eighth Route Army 's ;  
 7 the Eighth Route Army ;  
 7 the Eighth Route Army

3 This campaign ;  
 3 the Hundred Regiments Offensive ;

CHANGE ANSWERS

CONFIRM ANSWERS & PROCEED

# Annotation checks

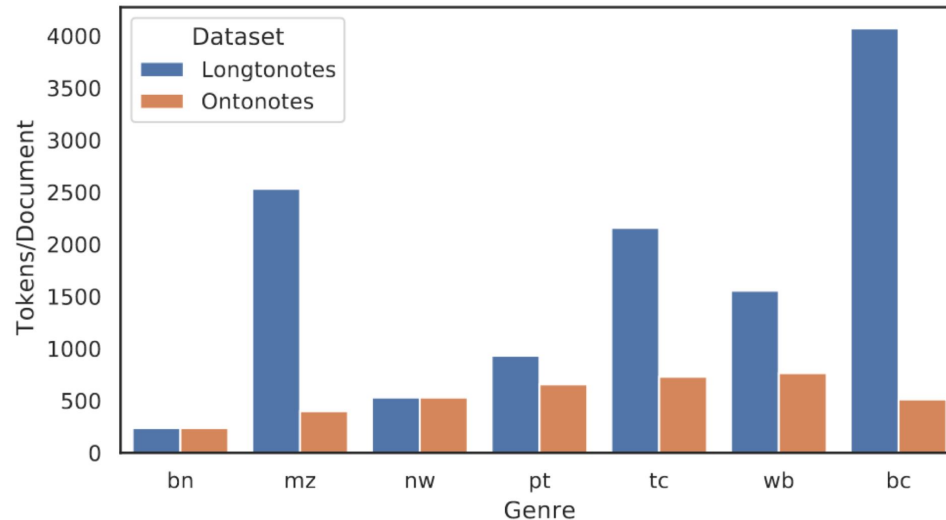
- 3 annotators were used to annotate the documents
  - University level English major
  - Annotators were paid 15 USD per hour
  - Their error analysis was done on 10% of the documents

# Annotation checks

- **3 annotators were used to annotate the documents**
  - University level English major
  - Annotators were paid 15 USD per hour
  - Their error analysis was done on 10% of the documents
    - **Over 90% agreement** (Krippendorff's Alpha and Strict matching)
    - Although disagreements increased with document length but very still manageable.

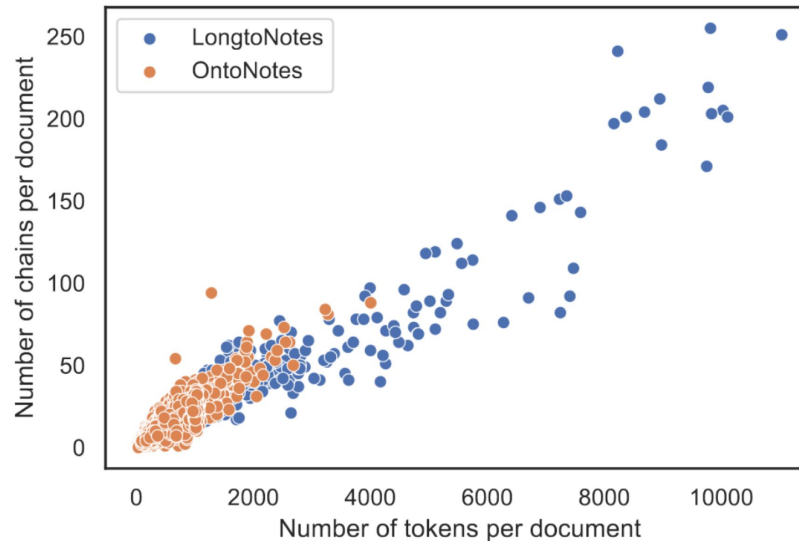
# LongtoNotes: More details

- Average document length -> **Upto 8X longer**



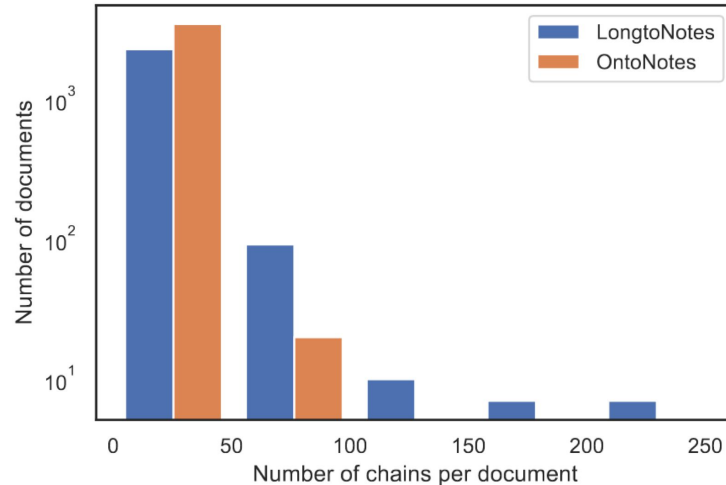
# LongtoNotes: More details

- Tokens per doc: **2615 (LongtoNotes)** vs 466 (OntoNotes)



# LongtoNotes: More details

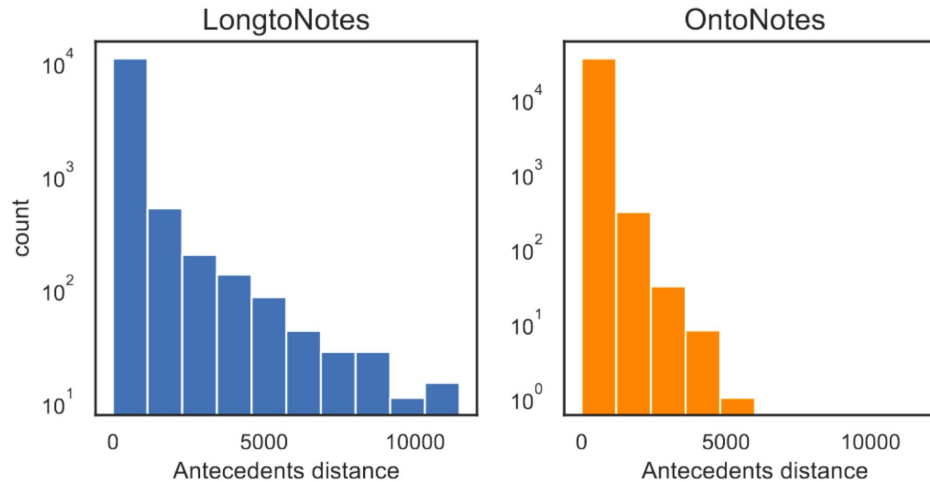
- Number of coreference chains per document -> **Up by 25%**





# LongtoNotes: More details

- Distance to antecedent -> **Upto 3X increased**



# Why should we care?

# Why should we care?

**State of the art** models (one year ago) **did not perform well** on long antecedent chains.

- Tested on 3 different models:
  - **Span based representation** (Higher order coref Joshi et al.)
  - **Token-wise representation** (Longformer Kirstain et al.)
  - **Memory networks** (Toshniwal et al.)

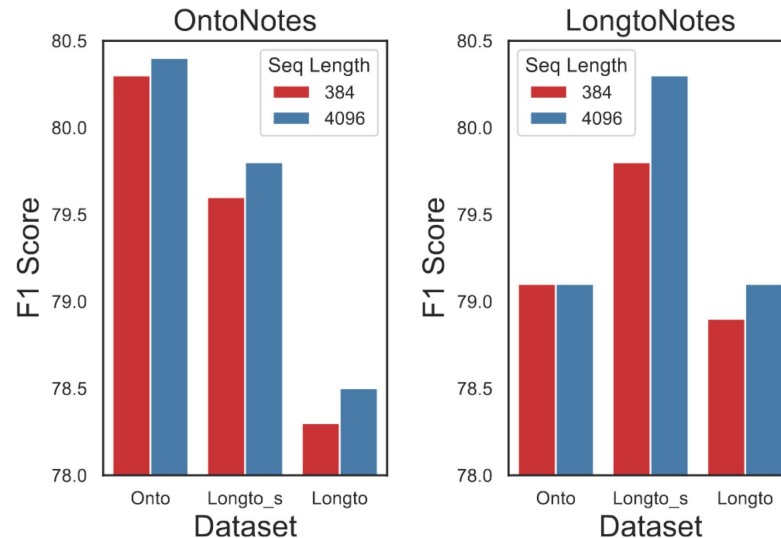
# Why should we care?

- **State of the art** models (one year ago) **did not perform well** on long antecedent chains.
- **Longer dependencies from LongtoNotes** can **assist** models in learning coreference better over longer chains.

# Chains	Training	SpanBERT	Token	Memory
≤ 40	Onto	<b>73.60</b>	<b>79.80</b>	<b>72.80</b>
	Longto	72.86	78.80	71.94
> 40	Onto	68.44	75.60	67.72
	Longto	<b>69.09</b>	<b>76.42</b>	<b>68.60</b>

# Why should we care?

- Longformer performance improved when trained on LongtoNotes.



# Summary

- We introduce **LongtoNotes**, a corpus of coreference-annotated documents of **significantly longer length** than what is currently available.
- **Can assist models in learning longer contexts** better when trained on LongtoNotes.

# Thank You

Email: [kmr.shridhr@gmail.com](mailto:kmr.shridhr@gmail.com)

Github:

<https://github.com/kumar-shridhar/LongtoNotes>

