Autoregressive Entity Retrieval (GENRE)

Noah Lee

Nicola De Cao, Gautier Izacard, Sebastian Riedel, Fabio Petroni
University of Amsterdam, Facebook AI Research ENS, PSL University,
Inria, University College London

Introduction

Entity Retrieval

Applications On…

- Recommender System
- Question Answering
- Chatbots

-> detect relevant vs irrelevant concepts
-> find relevant retrieval components

Introduction

Entity Retrieval

Search : John Smith...

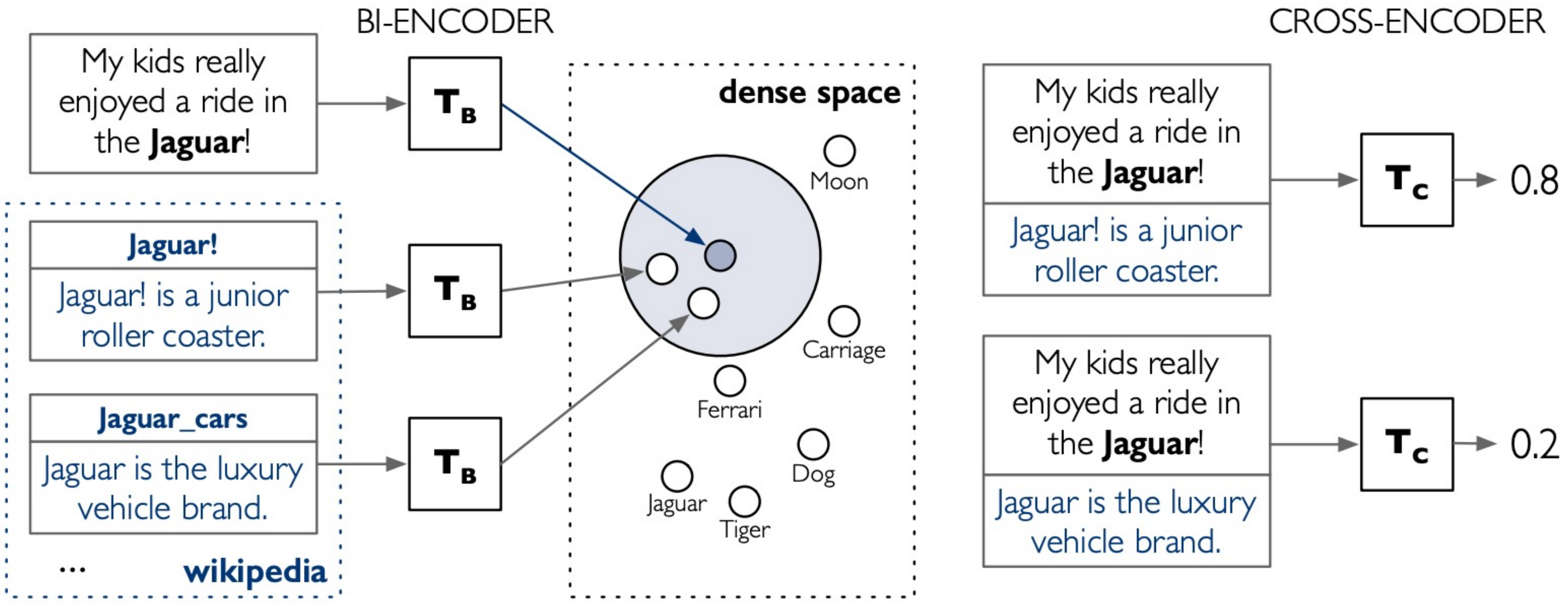
John Smith (actor)

John Smith (astronomer)

& many more...

Related Work

BLINK



Wu et al., 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval

Related Work

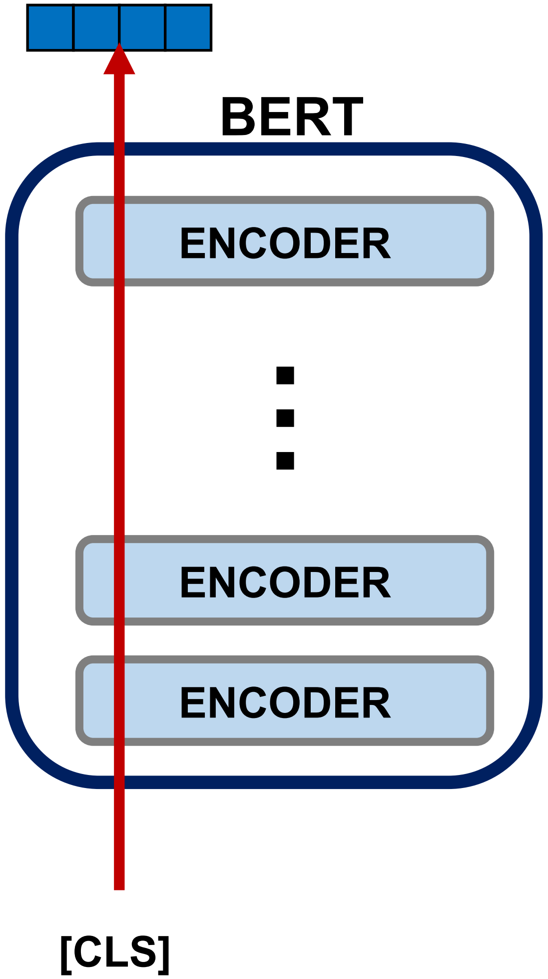
Dense Passage Retriever

- Bi-Encoder structure based on **BERT**

- Trainable Retriever (Pretrained Language Model)

- Uses the **output vector** corresponding to the **[CLS]** token

- Fine-tune to generate **dense embeddings**



Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

GENRE

Motivation

1. **Simple dot-product can miss fine-grained interactions between input & entity information**
   - Cross Encoder for re-ranking costly

2. **Large memory footprint needed to store dense vector embeddings (for MIPS)**
   - eg. ~24GB for 1024-D vectors for ~6M Wikipedia

3. **Arduous to subsample hard negative set @ training time**
   - Exact softmax over all entities very expensive

4. **Existing systems suffer from cold-start problem**
   - Entities with insufficient info

GENRE

Contribution

1. Propose GENRE, an autoregressive approach to directly capture relation between context & entity

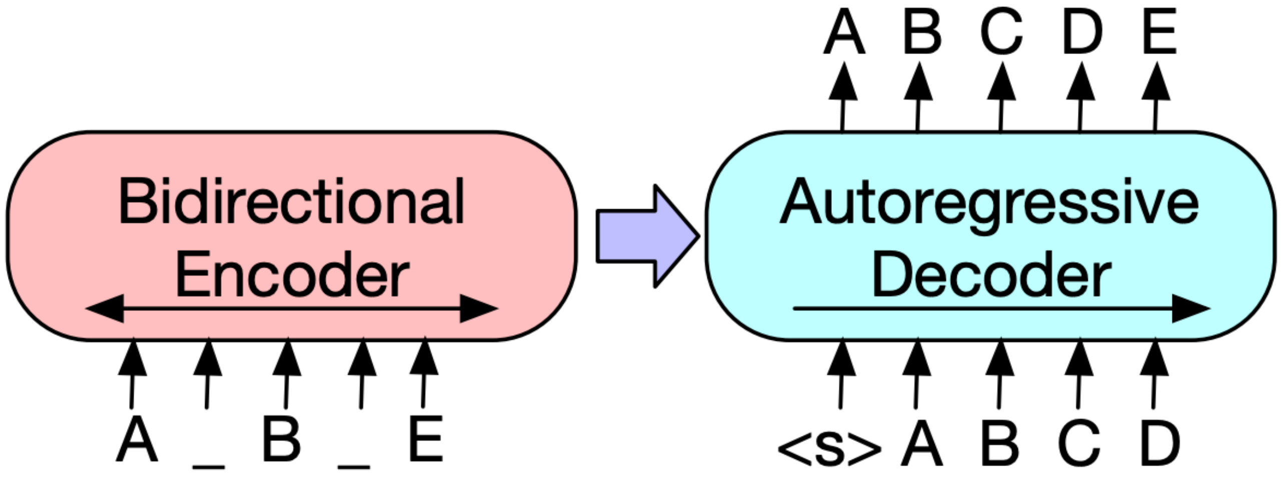
2. Memory footprint magnitudes smaller corresponding to the vocab size (not entity count)

3. Exact softmax computed without the need to subsample negative data

GENRE

Architecture

BART (large)



Pre-training
- Autoregressive denoising autoencoder

Fine-tuning
- Generating entity names

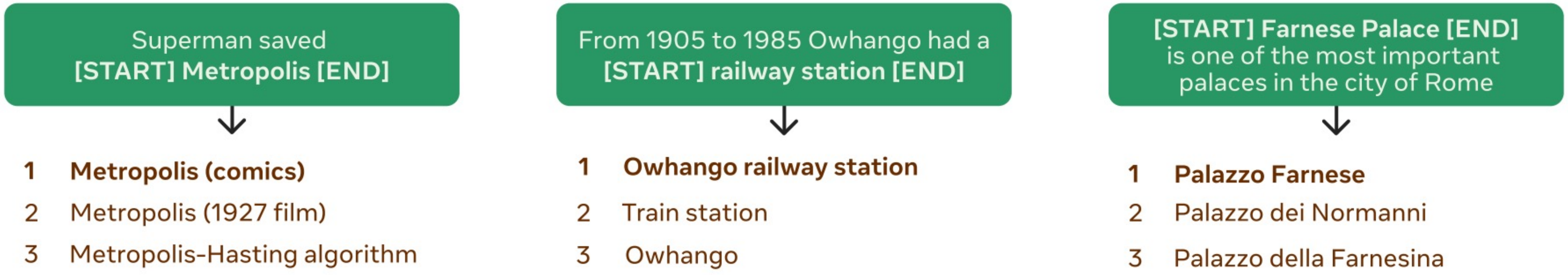
$$p_\theta(y|x) = \prod_{i=1}^{N} p_\theta(y_i|y_{<i}, x)$$

Inference
- Constrained Beam Search

GENRE

Examples

Entity Disambiguation

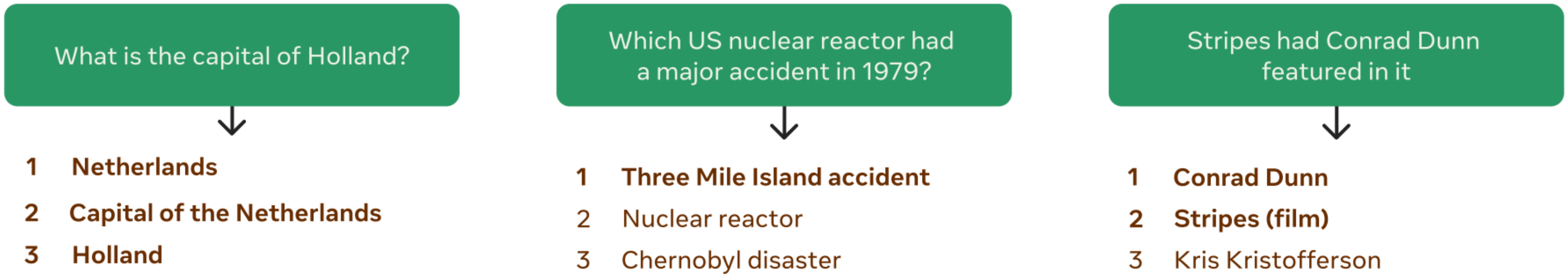


(a) Type specification. (b) Composing from context. (c) Translation.

GENRE

Examples

Document Retrieval



(d) Entity normalization. (e) Implicit factual knowledge. (f) Exact copy.

GENRE

Method

(Beam Search)

Objective : generate valid entity names… -> **PROBLEM**

Who created the World Wide Web? → History of

- WWW → • • •
- the Internet → • • •
- Web → • • •
- the web browser → • • •

GENRE

Method

Constrained Beam Search

Objective : generate valid entity names…

-> Constrain decoding search with a **prefix tree**

Who created the World Wide Web? → History of

~~WWW~~

the Internet → • • •

~~Web~~

the web browser → • • •

Experiments

Entity Disambiguation

```
ID: '87d95287-707e-4bd9-9633-ca0c611a4a3a_World_Without_Superma:8'
inputs: '[..] When Superman leaves Earth for New Krypton , he appoints , newly freed from the Phantom Zone , to take his place as guardian of [START_ENT] Metropolis [END_ENT] . Mon-El assumes the secret identity of Johnathan Kent as a tribute to Clark \'s adoptive father , posing as Clark \'s cousin . [..]'
gold_output: 'Metropolis (comics)'
predicted_outputs: [
    ('Metropolis_(comics)', -0.09),
    ('Themyscira_(DC_Comics)', -1.09),
    ('Metropolis_(disambiguation)', -1.27),
    ('Superman_(comic_book)', -1.51),
    ('Superman_(Earth-Two)', -1.52)
]
```

Experiments

Entity Disambiguation

Pre-train: BLINK dataset

	In-domain	Out-of-domain					
Method	**AIDA**	**MSNBC**	**AQUAINT**	**ACE2004**	**CWEB**	**WIKI***	**Avg.**
Ganea & Hofmann (2017)	92.2	93.7	88.5	88.5	77.9	77.5	86.4
Guo & Barbosa (2018)	89	92	87	88	77	84.5	86.2
Yang et al. (2018a)	**95.9**	92.6	89.9	88.5	**81.8**	79.2	88.0
Shahbazi et al. (2019)	93.5	92.3	90.1	88.7	78.4	79.8	87.1
Yang et al. (2019)	93.7	93.8	88.2	90.1	75.6	78.8	86.7
Le & Titov (2019)	89.6	92.2	**90.7**	88.1	78.2	81.7	86.8
Fang et al. (2019)	94.3	92.8	87.5	**91.2**	78.5	82.8	87.9
BLINK w/o candidate set**	79.6	80.0	80.3	82.5	64.2	75.5	77.0
GENRE	93.3	**94.3**	89.9	90.1	77.3	**87.4**	**88.8**

Experiments

E2E Entity Linking

Plain Text -> Hypertext

Michelle Obama has launched … as their White House challenger

→

<u>Michelle Obama</u> has launched … as their <u>White House</u> challenger

Experiments

E2E Entity Linking

```
ID: '1106testa_SOCCER'
inputs: 'SOCCER - RESULT IN SPANISH FIRST DIVISION. MADRID 1996-08-31 Result of game ↘
    played in the Spanish first division on Saturday: Deportivo Coruna 1 Real Madrid 1.'
gold_output: 'SOCCER - RESULT IN [SPANISH](Spain) FIRST DIVISION . [MADRID](Madrid) ↘
    1996-08-31 Result of game played in the [Spanish](Spain) first division on Saturday ↘
    : Deportivo Coruna 1 [Real Madrid](Real Madrid C.F.) 1.'
predicted_output: 'SOCCER - RESULT IN [SPANISH](Spain) FIRST DIVISION . [MADRID](Madrid) ↘
    1996-08-31 Result of game played in the [Spanish](Spain) first division on Saturday ↘
    : [Deportivo](Deportivo de La Coruna) Coruna 1 [Real Madrid](Real Madrid C.F.) 1.'
gold_spans: [
    [19, 7, 'Spain'],
    [44, 6, 'Madrid'],
    [91, 7, 'Spain'],
    [147, 11, 'Real_Madrid_C.F.']
]
predicted_spans: [
    [19, 7, 'Spain'],
    [44, 6, 'Madrid'],
    [91, 7, 'Spain'],
    [128, 9, 'Deportivo_de_La_Coruna'],
    [147, 11, 'Real_Madrid_C.F.']
]
```

Experiments

E2E Entity Linking

Pre-train: Abstract sections from Wikipedia

Method	In-domain AIDA	Out-of-domain MSNBC	Der	K50	R128	R500	OKE15*	OKE16*	Avg.
Hoffart et al. (2011)	72.8	65.1	32.6	55.4	46.4	**42.4**	**63.1**	0.0	47.2
Steinmetz & Sack (2013)	42.3	30.9	26.5	46.8	18.1	20.5	46.2	46.4	34.7
Moro et al. (2014)	48.5	39.7	29.8	<u>55.9</u>	23.0	29.1	41.9	37.7	38.2
Kolitsas et al. (2018)	<u>82.4</u>	<u>72.4</u>	34.1	35.2	**50.3**	38.2	<u>61.9</u>	<u>52.7</u>	53.4
Broscheit (2019)	79.3	-	-	-	-	-	-	-	
Martins et al. (2019)	81.9	-	-	-	-	-	-	-	
van Hulst et al. (2020)†	80.5	72.4	41.1	50.7	49.9	35.0	**63.1**	**58.3**	56.4
GENRE	**83.7**	**73.7**	**54.1**	**60.7**	46.7	<u>40.3</u>	56.1	50.0	**58.2**

Experiments

Page-level Document Retrieval

```
ID: 'sfq_18245'
inputs: "Which Florentine painter
    1535-1607 used the name Bronzino
    after the death of his 'uncle'?"
gold_output: 'Bronzino'
predicted_outputs: [
    ('Florence', -0.37),
    ('Bronzino', -0.62),
    ('Niccolo_Machiavelli', -0.64),
    ('Giorgio_de_Chirico', -0.71),
    ('Vitruvian_Man', -0.73)
]
```

(a) TriviaQA (open domain question answering).

Experiments

Page-level Document Retrieval

Train: KILT Dataset

	Fact Check.	Entity Disambiguation			Slot Filling		Open Domain QA				Dial.	
Model	**FEV**	**AY2**	**WnWi**	**WnCw**	**T-REx**	**zsRE**	**NQ**	**HoPo**	**TQA**	**ELI5**	**WoW**	**Avg.**
DPR + BERT	72.9	-	-	-	-	40.1	**60.7**	25.0	43.4	-	-	-
DPR	55.3	1.8	0.3	0.5	13.3	28.9	54.3	25.0	44.5	10.7	25.5	23.6
tf-idf	50.9	3.7	0.24	2.1	44.7	60.8	28.1	34.1	46.4	13.7	49.0	30.5
DPR + BART	55.3	75.5	45.2	46.9	13.3	28.9	54.3	25.0	44.4	10.7	25.4	38.6
RAG	61.9	72.6	48.1	47.6	28.7	53.7	59.5	30.6	48.7	11.0	57.8	47.3
BLINK + flair	63.7	81.5	80.2	68.8	59.6	78.8	24.5	46.1	65.6	9.3	38.2	56.0
GENRE	**83.6**	**89.9**	**87.4**	**71.2**	**79.4**	**95.8**	60.3	**51.3**	**69.2**	**15.8**	**62.9**	**69.7**

Further Analysis

Memory Footprint

Comparison between retrieval models on memory

Model	Memory	Param.	Index
DPR	70.9GB	220M	15B
RAG	40.4GB	626M	15B
BLINK	30.1GB	680M	6B
GENRE	**2.1GB**	**406M**	**17M**

1/14 BLINK
1/34 DPR

Uses entity names only

Further Analysis

Ablation Study

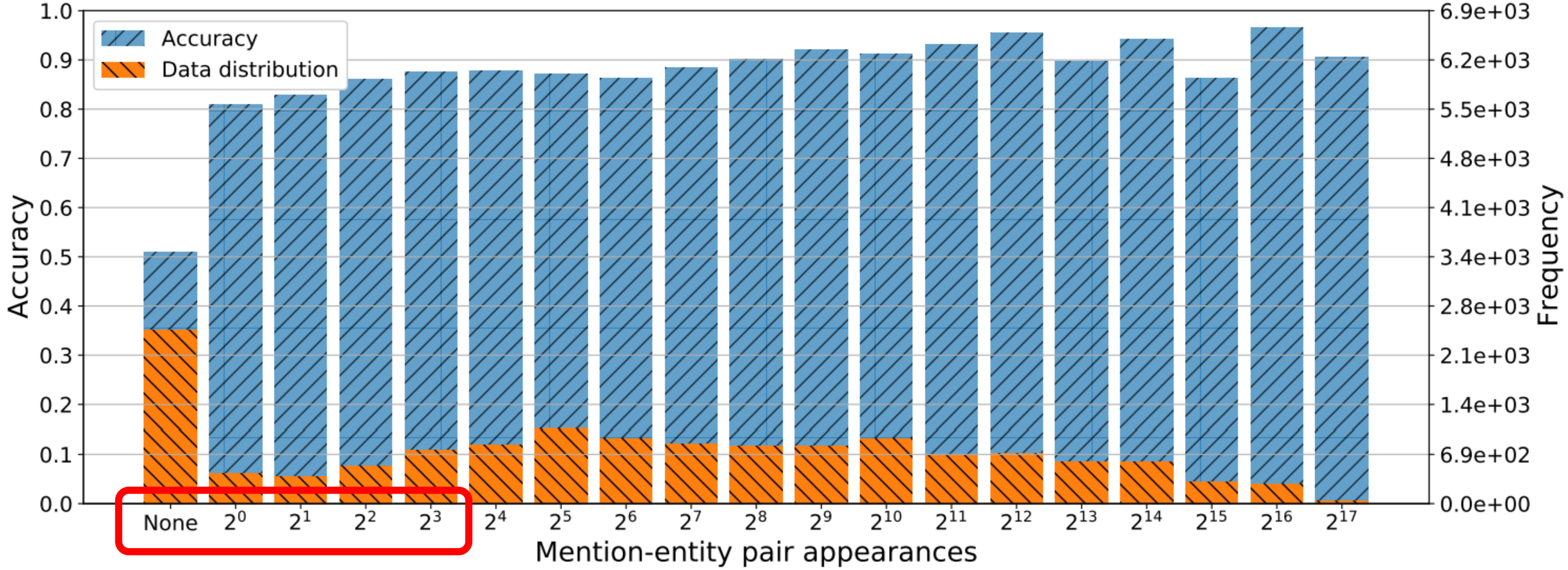
Ablation on Entity Disambiguation

Method	In-domain AIDA	MSNBC	Out-of-domain AQUAINT	ACE2004	CWEB	WIKI*	Avg.
GENRE	93.3	**94.3**	89.9	<u>90.1</u>	77.3	**87.4**	**88.8**
Ablations							
GENRE only AIDA data	88.6	88.1	77.1	82.3	71.9	71.7	80.0
GENRE only BLINK data	89.3	93.3	90.9	91.1	76.0	87.9	88.1
GENRE w/o candidate set	91.2	86.9	87.2	87.5	71.1	86.4	85.1
GENRE w/o constraints	86.4	80.0	81.7	82.1	66.0	81.1	79.6

Further Analysis

Entity Frequency

Accuracy per mention-entity pair frequency (on Entity Disambiguation)



Further Analysis

Cold Start

50 New Wikipedia Articles (2020)

- EM: 19/50 = 38%
- GENRE's bias on exactly copying mention

Type (support)	BLINK	GENRE	IDs*
Exact match (1543)	97.8	96.6	76.0
Partial match (1531)	70.7	86.9	63.8
No match (322)	49.4	59.9	55.0
Total (3396)	81.0	88.8	68.5

Conclusion

1. Propose GENRE, an autoregressive approach to directly capture relation between context & entity

2. Memory footprint magnitudes smaller corresponding to the vocab size (not entity count)

3. Exact softmax computed without the need to subsample negative data

Q&A