GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input

Tao Meng[1,2], Anjie Fang[2], Oleg Rokhlenko[2], Shervin Malmasi[2]

[1]University of California, Los Angeles [2]Amazon.com Inc.

Overview

Motivations

- Named Entity Recognition (NER) remains difficult in real-world settings.
- Existing benchmark datasets cannot represent these challenges.

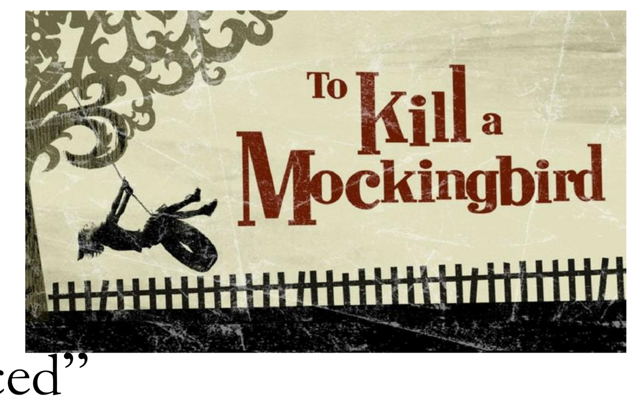
Contributions

- We developed new datasets to represent the current challenges in NER.
- We proposed GEMNET, a flexible architecture supporting external gazetteers.
- We analyzed the effects of gazetteer integration method, training method, gazetteer coverage and training data size

Current Challenges in NER

Short text:

- In voice, search, …
- e.g. search query *"<PROD>* reviews"



Long-tail entities:

- In domains with many entities
- e.g. *"A version for the* <u>*sega cd*</u> was announced"

Emerging entities:

- In domains with rapidly growing entities
- e.g. *"when did* <u>*AirPods Pro 2*</u> com out"

Complex entities:

- e.g. "what about <u>*To Kill a Mockingbird*</u>"

Newly Purposed Datasets

NER Taxonomy

- Follow WNUT 2017 taxonomy
- Emphasize complex entities

LOWNER (Low-Context Wikipedia NER)

- Source: Wikipedia
- Minimize the context around the entities
- e.g. "the regional capital is <u>*oranjestad, sint eustatius*</u>."

MSQ-NER (MS-MARCO Question NER)

- Source: MS-MARCO QnA corpus (V2.1)[1]
- Templatize the questions
- e.g. "where was *<CW>* filmed"

ORCAS-NER (Search Query NER)

- Source: ORCAS dataset[2]
- Templatize the user queries
- e.g. "*<PER>* parents"

Gazetteers

- Source: Wikidata KB
- Size: 1.67 million
- Follow our NER Taxonomy

Scan to get the datasets link

Set	Dataset	Type	# Sentence	# Token	# Entity	Avg. Sent Len	Entity Type Distribution PER	LOC	CORP	GRP	PROD	CW
1	LOWNER	Train	13,424	206,772	13,555	15.40±6.35	5,029	3,791	631	1,941	424	1,805
2	LOWNER	Dev	3,366	51,651	3,813	15.34±6.28	1,255	1,235	169	565	101	499
3	LOWNER	Test	1,385,290	21,303,399	490,749	15.37±6.29	215,411	120,480	20,015	52,566	15,976	74,830
4	MSQ-NER	Test	17,868	98,117	18,993	5.49±1.86	4,586	10,468	679	610	469	2,187
5	ORCAS-NER	Test	471,746	1,958,020	368,250	4.15±1.75	68,000	162,652	28,738	23,058	18,114	71,461

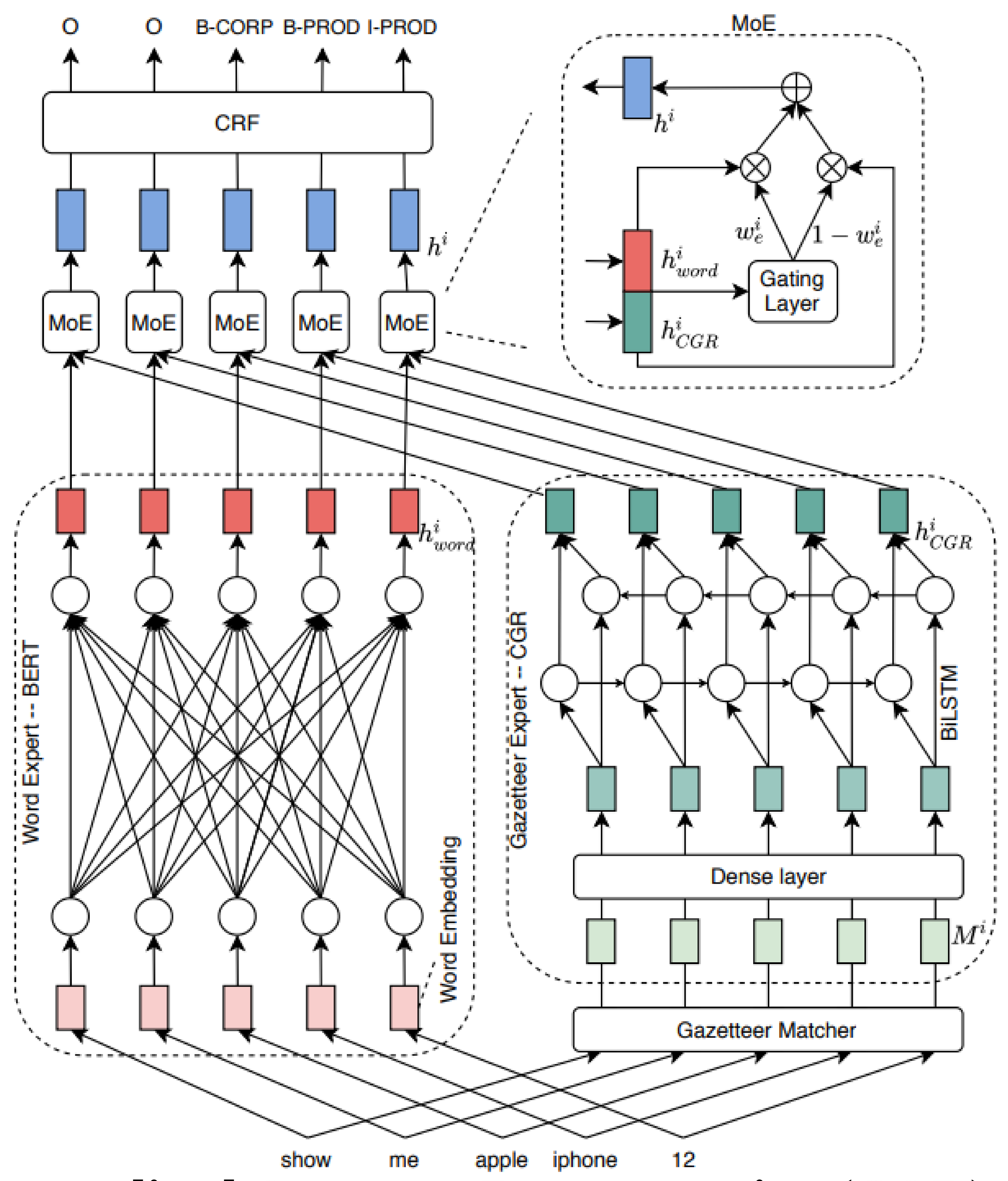
Benchmark models on our datasets

- Verify the challenges
- Poor context results in worse performance

CoNLL03	96.9	95.7	96.3
LOWNER	67.5	74.5	70.9
MSQ-NER	38.9	38.7	38.8
ORCAS-NER	56.8	51.6	54.1

[1] Bajaj, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268
[2]Craswell,et al. 2020. Orcas: 18 million clicked query-document pairs for analyzing search. arXiv preprint arXiv:2006.05324.

GEMNET Architecture



Contextualized Gazetteer Representation (CGR):

- Matching: A binary matrix
  - Easy to update

	O	B-PROD	I-PROD	B-CORP	I-CORP
How	1	0	0	0	0
much	1	0	0	0	0
is	1	0	0	0	0
Apple	0	1	0	1	0
iPhone	0	1	1	0	0
12	0	0	1	0	0

- Dense Embedding
- BiLSTM Encoding

Word Representation: BERT

Mixture of Experts (MoE):

- A gated architecture to conditional combine multiple experts.
- $w_e = \sigma(\theta[\mathbf{h}_{word}, \mathbf{h}_{CGR}])$,

$$\mathbf{h} = w_e \cdot \mathbf{h}_{word} + (1 - w_e) \cdot \mathbf{h}_{CGR},$$

Two-stage Training:

- Word expert is pre-trained but gazetteer expert is not.
- First stage: Freeze word expert
  Second stage: Joint training

Results

Benchmark Performance

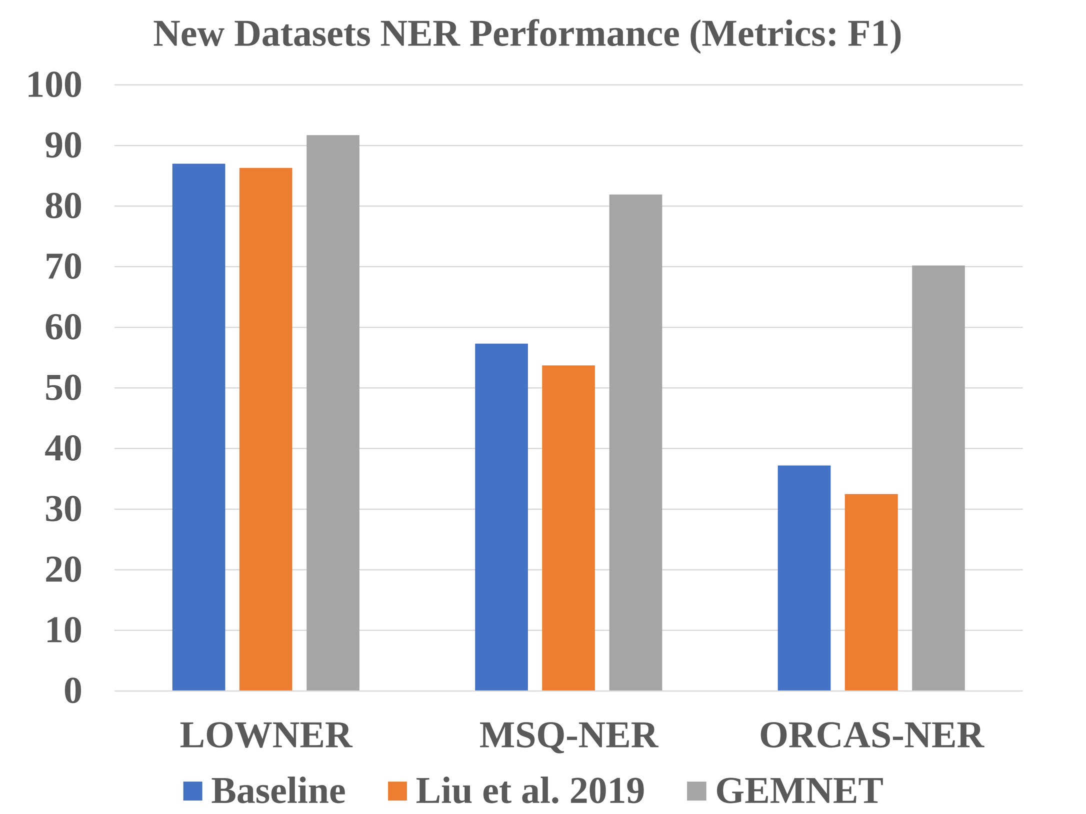
- Improves the baseline
- SOTA on WNUT17 uncased
- Demonstrate GEMNET is generally effective

Method	CoNLL03	WNUT17	OntoNotes
Uncased SOTA	91.0	46.1	88.1
BERT Baseline	89.6	46.9	86.9
GEMNET (BERT)	91.3	50.2	88.0

New Datasets Performance

- Improves the baseline remarkably
- Best architecture in comparison
- Improvements occurs especially for hard entities

Class	LOWNER	MSQ-NER	ORCAS-NER
PER	+1.9	+21.8	+40.1
LOC	+2.2	+37.5	+46.5
GRP	+8.5	+57.3	+57.2
CORP	+12.7	+57.7	+56.5
CW	+10.2	+58.8	+61.4
PROD	+10.7	+64.2	+62.0



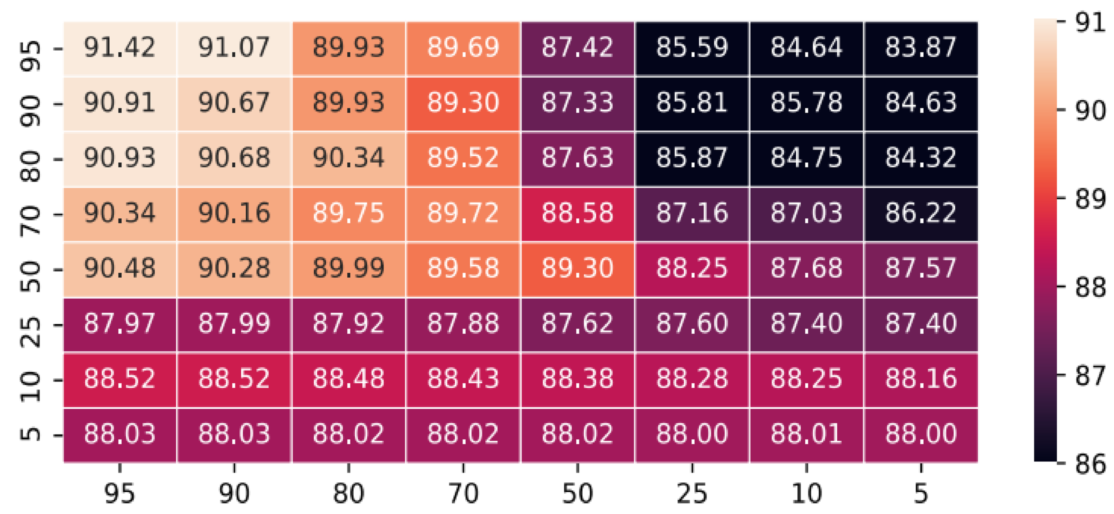
Ablation: Low-resource Setting

- Improves much faster with less data
- Achieves close to maximum with only 20% data

Size	LOWNER Baseline	LOWNER Ours	MSQ-NER Baseline	MSQ-NER Ours	ORCAS-NER Baseline	ORCAS-NER Ours
5%	74.2	81.2	52.2	60.4	33.1	44.2
10%	78.1	86.7	55.3	74.4	33.7	49.1
20%	81.2	88.7	55.5	81.3	33.9	69.7
100%	87.0	91.7	57.3	81.9	37.2	70.2

Ablation: Gazetteer Coverage

- By row: When the training coverage is fixed, Higher testing coverage → higher performance
- By column: When the testing coverage is fixed, closer coverage between training and testing → higher performance



Conclusion

- GEMNET supports external gazetteers, allowing the model's knowledge to be updated without retraining
- The datasets we released can serve as benchmarks for evaluating the entity knowledge possessed by models in future work
- To check our paper, please scan the QR code on the right.