Good Examples Makes A Fast Learner:

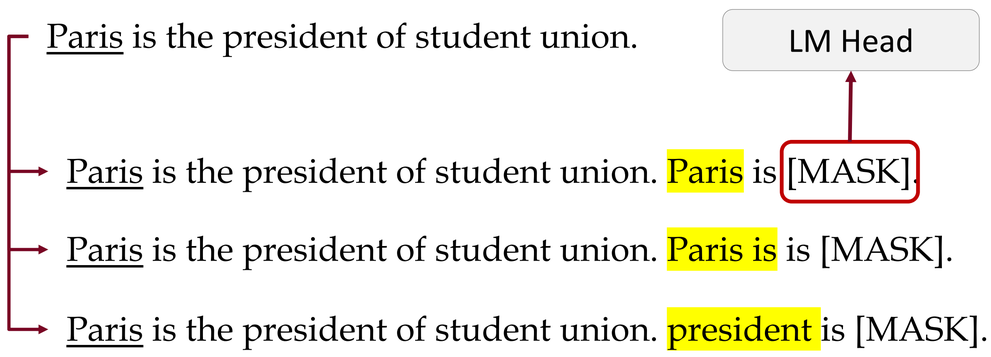
Simple Demonstration-based Learning for Low-resource NER

Dong-Ho Lee [1], Akshen Kadakia* [1], Kangmin Tan* [1], Mahak Agarwal [1], Xinyu Feng [1], Takashi Shibuya [2], Ryosuke Mitani [2], Toshiyuki Sekiya [2], Jay Pujara [1], Xiang Ren [1]

[1]University of Southern California, [2] R&D Center, Sony Group Corporation

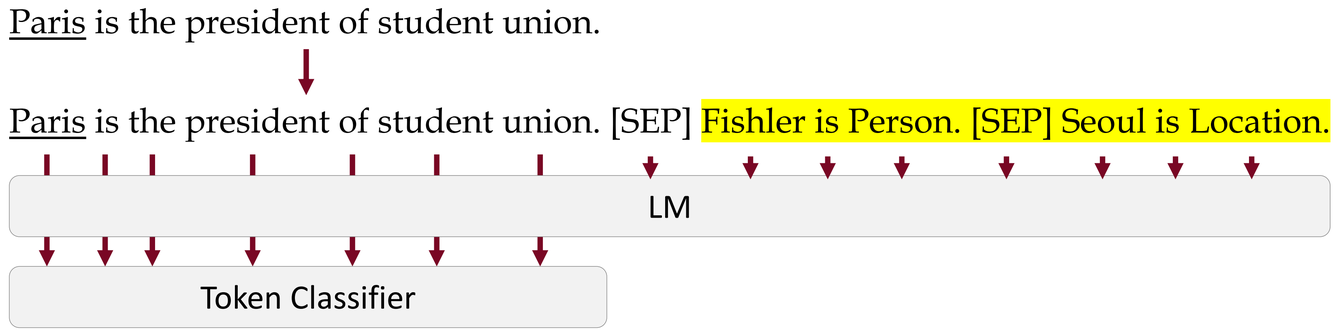
Motivation

Prompt-based Learning for NER (Cui et al., 2021)



1. Select entity candidate is expensive.
2. Neglect latent relationship among token
3. Cannot be applied to existing token classification module.

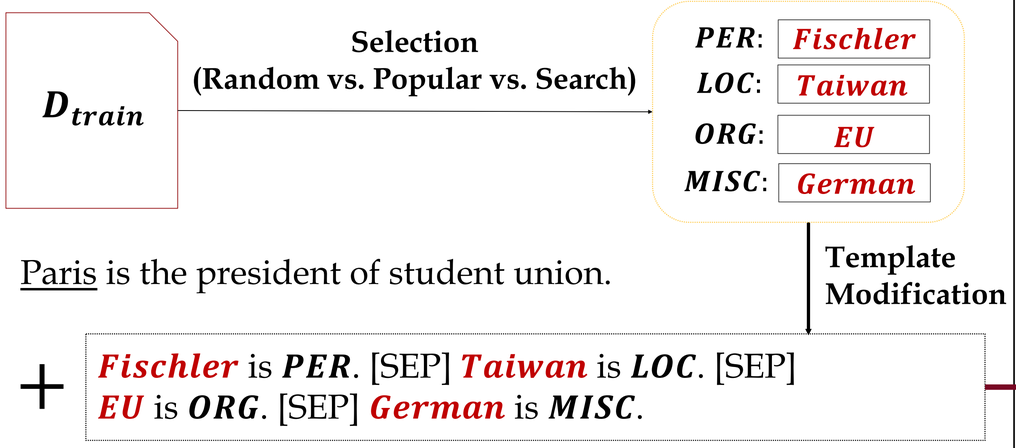
Ours: Demonstration-based Learning for NER



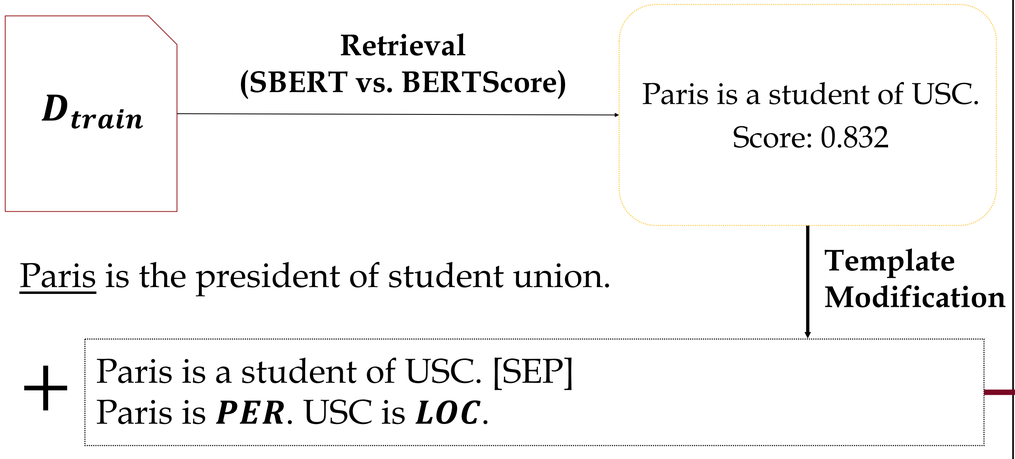
Simple additional task demonstration as a context can be helpful and cost-effective !

Task Demonstrations

(a) Entity-Oriented Demonstration



(b) Instance-Oriented Demonstration



Selection vs. Retrieval

Templates

(a) Entity-Oriented Demonstration

no-context	***Fischler*** is ***PER***. [SEP] ***Taiwan*** is ***LOC***. [SEP] ***EU*** is ***ORG***. [SEP] ***German*** is ***MISC***.
context	France backed ***Fischler***'s proposal. ***Fischler*** is ***PER***. [SEP] She has complained of back pain since her trip to ***Taiwan***. ***Taiwan*** is ***LOC***. [SEP] EU rejects German call. ***EU*** is ***ORG***. [SEP] EU rejects German call. ***German*** is ***MISC***.
lexical	France backed ***PER*** proposal. [SEP] She has complained of back pain since her trip to ***LOC***. [SEP] ***ORG*** rejects German call. [SEP] EU rejects ***MISC*** call.

(b) Instance-Oriented Demonstration

context	***Paris*** is a student of ***USC***. [SEP] ***Paris*** is ***PER***. ***USC*** is ***ORG***.
lexical	***PER*** is a student of ***ORG***.

Train Process

(1) Create Task Demonstration: $\tilde{x} = [[SEP]; \widehat{x_1}; \ldots; \widehat{x_n}]$

(2) Concatenate input and demonstration: $[x; \tilde{x}]$

(3) Feed the concatenated input to Transformer embedder: $[h; \tilde{h}] = embed([x; \tilde{x}])$

(4) Feed the embedding of original input to the token classification layer: $p_\theta(y|h)$

(5) Minimize the conditional probability by cross entropy: $-\sum_{i=1}^{n} \log p_\theta(y|h)$

In-domain Performance

Demonstration / Method	Strategy	Template	CoNLL03		Ontonotes 5.0		BC5CDR	
			25	50	25	50	25	50
BERT+CRF w/o demonstration	-	-	52.72 ±2.44	62.75 ±0.98	38.97 ±4.62	54.51 ±3.27	52.56 ±0.46	60.20 ±2.01
BERT+CRF w/ Instance-oriented demonstration	SBERT (variable)	lexical	48.92 ±2.81	57.68 ±0.37	36.58 ±4.61	44.47 ±2.58	49.41 ±0.94	51.98 ±2.14
		context	53.62 ±1.64	64.21 ±1.87	42.18 ±5.21	53.07 ±3.46	54.71 ±2.09	59.78 ±1.47
	BERTScore (variable)	lexical	49.55 ±3.18	58.85 ±1.06	35.42 ±3.88	44.70 ±2.41	49.37 ±0.19	51.61 ±2.45
		context	53.97 ±1.52	64.66 ±2.04	37.56 ±5.29	53.13 ±3.22	54.81 ±2.11	59.63 ±1.94
BERT+CRF w/ Entity-oriented demonstration	random (variable)	no-context	53.95 ±1.89	63.31 ±2.14	42.25 ±3.61	55.71 ±3.82	53.58 ±0.48	59.97 ±1.89
		lexical	55.20 ±2.24	63.60 ±2.32	44.02 ±4.73	56.31 ±3.83	53.79 ±0.61	59.65 ±1.71
		context	54.84 ±2.12	63.51 ±2.83	43.57 ±3.73	56.76 ±3.69	54.08 ±0.97	59.94 ±1.70
	popular (fixed)	no-context	54.34 ±3.33	64.30 ±2.76	43.02 ±4.33	56.65 ±3.35	53.86 ±0.86	60.51 ±1.77
		lexical	56.22 ±3.88	64.95 ±2.04	45.31 ±5.02	58.24 ±3.17	54.14 ±0.67	60.67 ±1.58
		context	56.52 ±3.34	64.47 ±2.35	45.52 ±4.69	58.40 ±3.24	54.31 ±0.80	61.31 ±1.51
	search (fixed)	no-context	54.63 ±2.12	64.50 ±2.76	42.88 ±5.41	56.96 ±4.09	53.97 ±1.32	60.84 ±2.14
		lexical	56.57 ±3.61	**65.11 ±2.71**	44.87 ±5.09	58.51 ±3.42	54.39 ±1.57	60.76 ±2.12
		context	**57.00 ±4.03**	64.82 ±3.16	**45.74 ±5.57**	**59.00 ±3.27**	**55.83 ±1.25**	**62.87 ±2.41**

1. Entity-oriented demonstration is better than instance-oriented demonstration.
2. Adding context improves examples (context, lexical > no-context).

Domain Adaptation Performance

Baselines	*Label Sharing*		*Label Different*	
	CoNLL03 -> Ontonotes		CoNLL03 -> BC5CDR	
	25	50	25	50
BERT+CRF w/o demonstration	61.22 ±1.93	66.44 ±1.75	52.31 ±1.02	62.10 ±1.01
NNShot	46.67 ±5.48	46.34 ±2.66	44.93 ±1.78	48.12 ±2.72
StructShot	43.61 ±4.58	43.02 ±3.19	25.86 ±4.14	27.81 ±2.10

Strategy	Template				
popular (fixed)	no-context	62.31 ±1.60	69.39 ±1.59	54.33 ±0.80	62.87 ±0.23
	lexical	62.50 ±2.41	69.34 ±1.38	54.30 ±1.12	63.05 ±0.45
	context	62.59 ±2.38	69.91 ±1.24	54.45 ±0.96	63.40 ±0.33
search (fixed)	no-context	62.38 ±2.47	69.57 ±1.50	54.51 ±2.25	62.93 ±1.96
	lexical	62.51 ±2.43	68.93 ±1.69	54.70 ±2.26	62.88 ±2.90
	context	62.63 ±2.94	**69.98 ±1.63**	**54.97 ±1.99**	**63.55 ±1.58**

1. Transfer the embedder weights from source to target model = **Task-adaptive pre-training** on NER task formats.
2. Demonstration-based learning **allows the source model to adapt the target domain quickly.**