Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning

ICLR 2023

Sheng Zhang, Hao Cheng, Jianfeng Gao, Hoifung Poon

발제자: 윤예준

https://arxiv.org/pdf/2208.14565.pdf

01. 연구배경

Named entity recognition(NER): Identifying text spans associated with proper names and classifying them into a predefined set of semantic classes such as person, location, organization

- 이전 연구는 주로 NER을 sequence labeling or span classification로 접근함

- 대조 학습을 사용한 기존 NER 연구들은 모든 non-entity tokens/spans를 Outside(O)와 같은 동일한 클래스로 지정하고 있기 때문에 annotation이 잘못된 경우 false negative noises를 발생시킬 수 있음.
  (Few-Shot Named Entity Recognition via Contrastive Learning, ACL 2022)

open-domain question answering의 최근 성공에 영감을 받아 대조 학습을 적용하는 NER을 위한 효율적인 bi-encoder framework(BINDER) 제안
(Dense Passage Retrieval for Open-Domain Question Answering, ACL 2020,
간단한 dual-encoder framework를 통해 적은 수의 questions과 passages로 학습된 dense representations만을 사용하여 retrieval을 구현할 수 있음을 보여줌)

02. 제안 방법

1. Bi-Encoder for NER
   - Entity Type Encoder와 Text Encoder로 이루어짐 (BERT 사용)
   - NER task를 위해선 2개의 input 고려
     - entity type descriptions
     - text to detect named entities
   - Entity Type Encoder는 entity에 대한 representations 출력
   - Text Encoder는 named entities가 잠재적으로 언급되는 주어진 텍스트의 각 토큰에 대한 representations 출력

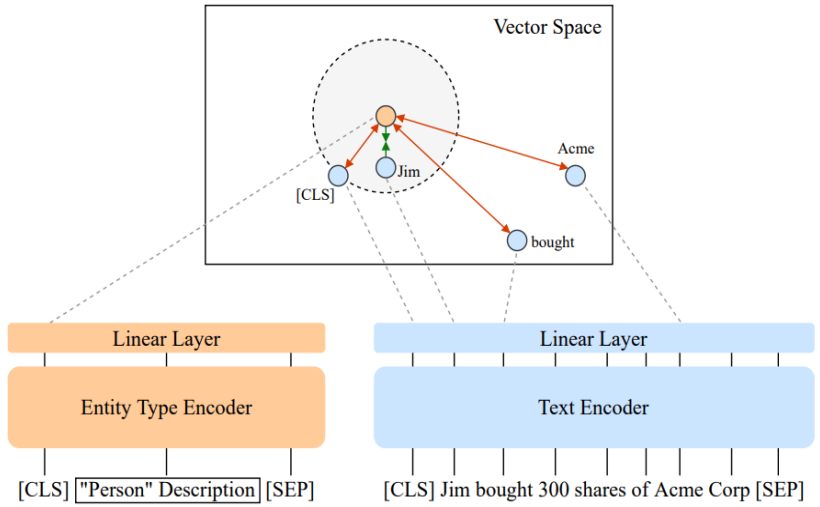
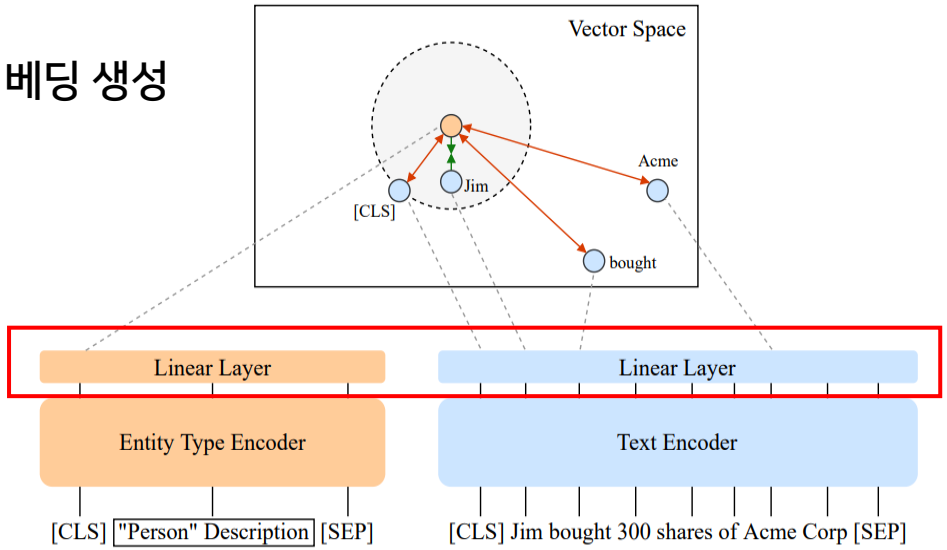


Figure 1: The overall architecture of **BINDER**. The entity type and text encoder are isomorphic and fully decoupled Transformer models. In the vector space, the anchor point (○) represents the special token `[CLS]` from the entity type encoder. Through contrastive learning, we maximize the similarity between the anchor and the positive token (○`Jim`), and minimize the similarity between the anchor and negative tokens. The dotted gray circle (delimited by the similarity between the anchor and ○`[CLS]` from the text encoder) represents a threshold that separates entity tokens from non-entity tokens. To reduce clutter, we do not draw data points that represent possible spans from the input text.

02. 제안 방법

Bi-Encoder for NER

1. Entity Type Embeddings
   - 대조 학습을 위해 벡터 공간에서 anchors 역할을 하는 entity type 임베딩 생성
   - $BERT^E$: Entity Type Encoder
   - E_k: k번째 entity type description $\mathcal{E} = \{E_1, \ldots, E_K\}$
   - Set of entity types:
     (each entity type has one or multiple natural language descriptions)



$$\mathbf{h}^{E_k}_{[\mathrm{CLS}]} = \mathrm{BERT}^E(E_k), \quad (1)$$

$$\mathbf{e}_k = \mathrm{Linear}^E(\mathbf{h}^{E_k}_{[\mathrm{CLS}]}), \quad (2)$$

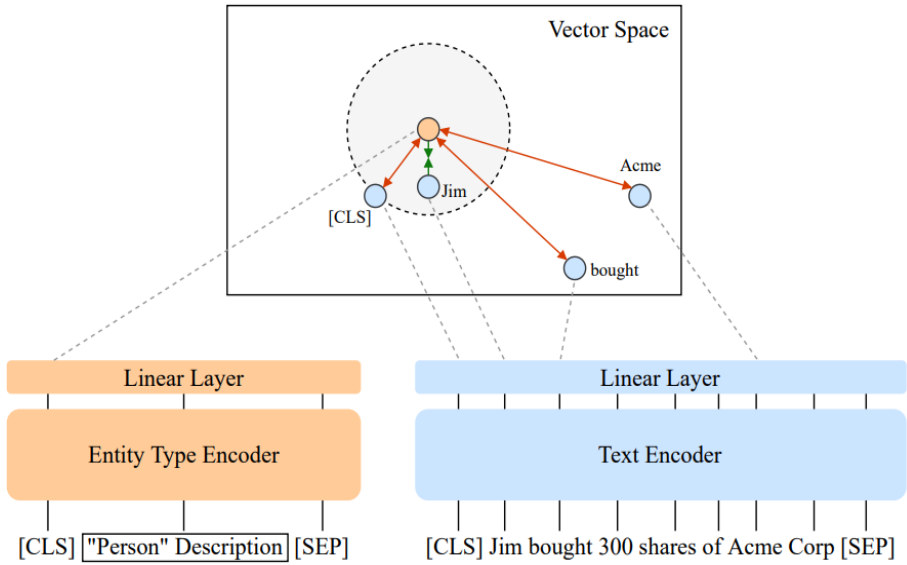
2. Text Token Embeddings
   - $BERT^T$: Text Encoder
   - [CLS] 임베딩 대신 entity span embeddings과의 유사성을 계산하기 위한 기본 단위로 text token embeddings을 사용하는 것을 고려
   - 입력 텍스트에 사전에 알려지지 않은 multiple potential NER이 있기 때문에 그냥 사용하면 NER 작업에 막대한 계산 오버헤드 발생할 수 있음

$$\mathbf{h}^T_1, \ldots, \mathbf{h}^T_N = \mathrm{BERT}^T(x_1, \ldots, x_N), \quad (3)$$

$$\mathbf{p}_n = \mathrm{Linear}^T(\mathbf{h}^T_n), \quad (4)$$

02. 제안 방법

2. NER Contrastive Learning
   - Entity Type Embeddings와 Text Token Embeddings을 기반으로 두가지 대조 방법 이용
   - span(i, j): i 위치에 start token, j 위치에 end token을 가진 입력 텍스트의 contiguous token sequence라고 가정
   - 전반적인 목표는 entity가 언급된 span representations을 해당 entity type의 embeddings(positive)에 가깝게 관련 없는 type과는 멀리 밀어내는 것
   - 이를 위해 span과 token embedding space를 기반으로 하는 두 가지 objectives를 제안



02. 제안 방법

NER Contrastive Learning

1. Span-based Objective
   - span (i, j)에 대한 vector representations를 위한 방법 고려
   - $\oplus$: vector concatenation
   - $D(j-i) \in R^{L \times m}$: row from the span width embedding matrix
     L: maximum span width considered

$$\mathbf{s}_{i,j} = \texttt{Linear}^{S}(\mathbf{h}_i^T \oplus \mathbf{h}_j^T \oplus D(j-i)), \quad (5)$$

$$\ell_{\text{span}} = -\log \frac{\exp(\text{sim}(\mathbf{s}_{i,j}, \mathbf{e}_k))}{\sum_{\mathbf{s}' \in \mathcal{S}_k^- \cup \mathbf{s}_{i,j}} \exp(\text{sim}(\mathbf{s}', \mathbf{e}_k))}, \quad (6)$$

2. Position-based Objective
   - span-based objective는 한가지 한계점이 존재:
     부분적으로 올바른 span이든 gold entity span과
     겹치지 않는 span이든 동일한 방식으로 계산됨

   - Linear layer를 추가 도입하여 시작과 끝 위치에 대한
     계산 도입

   - 식 (6)과 주요 차이점은 시작과 끝의 위치가 서로 독립적인
     해당 negative sets에서 비롯되기 때문에 entity의 시작 및
     끝 위치를 예측하는데 잠재적으로 도움이 될 수 있음.

$$e_k^{\text{B}} = \texttt{Linear}_B^E(\mathbf{h}_{[\texttt{CLS}]}^{E_k}) \quad (7)$$

$$e_k^{\text{Q}} = \texttt{Linear}_Q^E(\mathbf{h}_{[\texttt{CLS}]}^{E_k}), \quad (8)$$

$$\mathbf{u}_n = \texttt{Linear}_B^T(\mathbf{h}_n^T), \quad (9)$$

$$\mathbf{v}_n = \texttt{Linear}_Q^T(\mathbf{h}_n^T). \quad (10)$$

$$\ell_{\text{start}} = -\log \frac{\exp(\text{sim}(\mathbf{u}_i, \mathbf{e}_k^{\text{B}}))}{\sum_{\mathbf{u}' \in \mathcal{U}_k^- \cup \mathbf{u}_i} \exp(\text{sim}(\mathbf{u}', \mathbf{e}_k^{\text{B}}))} \quad (11)$$

$$\ell_{\text{end}} = -\log \frac{\exp(\text{sim}(\mathbf{v}_j, \mathbf{e}_k^{\text{Q}}))}{\sum_{\mathbf{v}' \in \mathcal{V}_k^- \cup \mathbf{v}_j} \exp(\text{sim}(\mathbf{v}', \mathbf{e}_k^{\text{Q}}))}, \quad (12)$$

02. 제안 방법

NER Contrastive Learning

3. Thresholding for Non-Entity Cases
   - 모델이 positive로 간주하기 전에 span이 얼마나 유사해야 하는지 결정하는 것은 문제가 될 수 있음.
   - 즉, entity span과 non-entity span을 명확하게 분리할 수 없음
   - 이 문제를 해결하기 위해 [CLS] 토큰과 entity 유형 간의 유사성을 dynamic threshold으로 사용

$$\ell^{+}_{start} = \beta\ell_{start} - (1-\beta)log\frac{\exp(sim(u_{[cls]}, e^{B}_{k}))}{\sum_{u'\in\mathcal{U}^{-}_{k}}\exp(sim(u', e^{B}_{k}))}$$

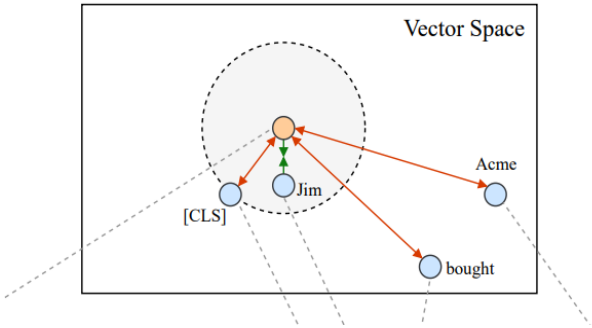
$$\ell^{+}_{end} = \beta\ell_{end} - (1-\beta)log\frac{\exp(sim(v_{[cls]}, e^{Q}_{k}))}{\sum_{v'\in\mathcal{V}^{-}_{k}}\exp(sim(v', e^{Q}_{k}))}$$

$$\ell^{+}_{span} = \beta\ell_{span} - (1-\beta)log\frac{\exp(sim(s_{0,0}, e_{k}))}{\sum_{s'\in\mathcal{S}^{-}_{k}\cup s_{i,j}}\exp(sim(s', e_{k}))}$$

4. 최종 학습 손실 함수

$$\mathcal{L} = \alpha\ell^{+}_{\text{start}} + \gamma\ell^{+}_{\text{end}} + \lambda\ell^{+}_{\text{joint}}, \qquad (15)$$

where α, γ, λ are all scalar hyperparameters.



02. 제안 방법

NER Contrastive Learning

3. Inference Strategy
   - joint position-span prediction
     - 학습된 null threshold보다 낮은 similar 점수를 가진 start or end의 span(i,j)를 제거
   - span-only prediction
     - span null threshold보다 높은 span similarity 점수는 positive로 예측

Algorithm 1: Inference for BINDER.

Input: $\mathcal{S} = \{(i,j) | i,j = 1,\ldots,N, 0 \leq j-i \leq L\}$ the set of spans , $\mathcal{E} = \{E_1,\ldots,E_K\}$ the set of entity types, `joint` for whether using joint position-span inference, and `flat` for whether the inference is for flat NER.

Function `main()`:
- $M = \{\}$;
- D = Dict() # a dictionary maps item in M to its similarity score;
- **for** $E_k \in \mathcal{E}$ **do**
  - calculate the threshold scores
    - $b_{null} = \text{sim}(\mathbf{u}_{\texttt{[CLS]}}, \mathbf{e}^{\text{B}}_k)$,
    - $e_{null} = \text{sim}(\mathbf{v}_{\texttt{[CLS]}}, \mathbf{e}^{\text{Q}}_k)$,
    - $s_{null} = \text{sim}(\mathbf{s}_{0,0}, \mathbf{e}_k)$;
  - **for** $(i,j) \in \mathcal{S}$ **do**
    - calculate the similarity scores
      - $b = \text{sim}(\mathbf{u}_i, \mathbf{e}^{\text{B}}_k)$,
      - $e = \text{sim}(\mathbf{v}_j, \mathbf{e}^{\text{Q}}_k)$,
      - $s = \text{sim}(\mathbf{s}_{i,j}, \mathbf{e}_k)$;
    - **if** *`joint` is true and* $b < b_{null}$ *or* $e < e_{null}$ **then**
      - Continue;
    - **end**
    - **if** $s > s_{null}$ **then**
      - $M = M \bigcup \{(i,j,E_k)\}$;
      - $D[(i,j,E_k)] = s$;
    - **end**
  - **end**
- **end**
- **if** *flat is true* **then**
  - **return** `removeOverlap(`D`)`;
- **end**
- **return** M;

Function `removeOverlap(`$\hat{D}$`)`:
- $\hat{M} = \{\}$;
- sort $\hat{D}$ by the similarity score in descending order and break the tie by ascending in start and end positions;
- **for** (i,j,E_k) *in* $\hat{D}$ **do**
  - **if** *span* (i,j) *has no overlap in* $\hat{M}$ **then**
    - $\hat{M} = \hat{M} \cup \{(i,j,E_k)\}$;
  - **end**
- **end**
- **return** $\hat{M}$;

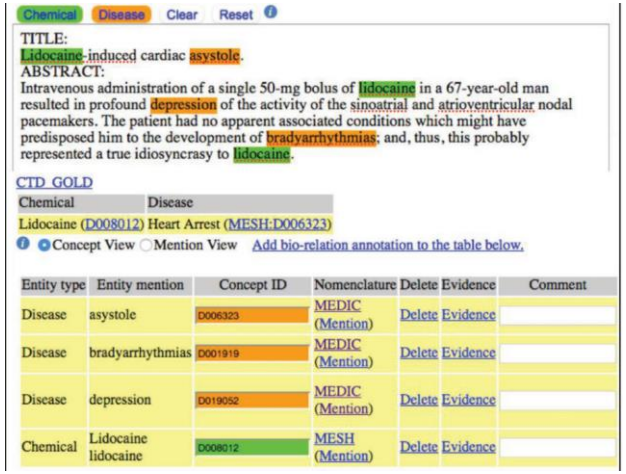
03. 실험 결과

Evaluation Metrics
- micro F1-Score
- predicted entity span은 span 경계와 predicted entity type이 모두 정확한 경우 올바른 것으로 간주

Datasets
- Nested NER: ACE2004, ACE2005, and GENIA
  (ACE2004 and ACE2005 are collected from general domains)

- Flat: BC5-chem/disease, NCBI, BC2GM, and JNLPBA
  (biomedical NER datasets)


```
"sentence": "Earlier documents in the case have included embarrassing details about perks Welch received
"golden-entity-mentions": [
  {
    "text": "Welch",
    "entity-type": "PER:Individual",
    "head": {
      "text": "Welch",
      "start": 11,
      "end": 12
    },
    "entity_id": "APW_ENG_20030325.0786-E24-38",
    "start": 11,
    "end": 12
  },
```

03. 실험 결과

	Encoder	ACE2004			ACE2005			GENIA		
		P	R	F1	P	R	F1	P	R	F1
Lu and Roth (2015)	-	70.0	56.9	62.8	66.3	59.2	62.5	74.2	66.7	70.3
Katiyar and Cardie (2018)	L	73.6	71.8	72.7	70.6	70.4	70.5	77.7	71.8	74.6
Shibuya and Hovy (2020)	Bl	83.7	81.9	82.8	83.0	82.4	82.7	78.1	76.5	77.3
Wang et al. (2020)	Bl/BioB	86.1	86.5	86.3	84.0	85.4	84.7	79.5	78.9	79.2
Li et al. (2020)†	Bl/BioB	85.8	85.8	85.8	85.0	84.1	84.6	81.2	76.4	78.7
Yu et al. (2020)	Bl/BioB	87.3	86.0	86.7	85.2	85.6	85.4	81.8	79.3	**80.5**
Tan et al. (2021)	Bl/BioB	88.5	86.1	<u>87.3</u>	87.5	86.6	<u>87.1</u>	82.3	78.7	<u>80.4</u>
Yan et al. (2021)	BAl	87.3	86.4	86.8	83.2	86.4	84.7	78.6	79.3	78.9
Zhang et al. (2022)	T5b	86.5	84.5	85.4	83.3	86.6	84.9	81.0	77.2	79.1
Wan et al. (2022)	Bb	86.7	85.9	86.3	84.4	85.9	85.1	77.9	80.7	79.3
Ours	Bb/BioB	88.3	89.1	**88.7**	89.1	89.8	**89.5**	81.5	79.6	**80.5**

Table 1: Test scores on three nested NER datasets. Bold and underline indicate the best and the second best respectively. The different encoders are used: L = LSTM, Bl = BERT-large, BioB = BioBERT, BAl = BART-large, T5b = T5-base, Bb = BERT-base. † Original scores are not reproducible. We report the rerun of their code. Similar scores are also reported in Yan et al. (2021).

03. 실험 결과

BLURB is the Biomedical Language Understanding and Reasoning Benchmark.

	Encoder	BC5-chem	BC5-disease	NCBI	BC2GM	JNLPBA
Lee et al. (2019)	BioBERT	92.9	84.7	89.1	83.8	78.6
Gu et al. (2021)	PubMedBERT	93.3	85.6	87.8	84.5	79.1
Kanakarajan et al. (2021)	BioELECTRA	93.6	85.8	<u>89.4</u>	<u>84.7</u>	<u>80.2</u>
Yasunaga et al. (2022)	LinkBERT	<u>93.8</u>	<u>86.1</u>	88.2	**84.9**	79.0
Ours	PubMedBERT	**95.0**	**88.0**	**90.9**	84.6	**80.3**

Table 3: Test F1 scores on five flat NER datasets from the BLURB benchmark (aka.ms/blurb). Bold and underline indicate the best and the second best respectively. All encoders use their base version.

04. 결 론

- text와 entity types을 같은 벡터 공간에 별도로 매핑하는 NER용 bi-encoder framework 제시

- NER을 metric 학습 문제로 공식화함으로써, entity span 식별과 분류를 동시에 학습하기 위해 새로운 대조 손실 사용을 제안

- 향후 self-supervised or zero-shot settings에서 소개한 방법 적용 예정.

감사합니다.

A.

	BC5CDR		
	P	R	F1
Distantly Supervised			
Dict/KB Matching	86.4	51.2	64.3
AutoNER (Shang et al., 2018)	82.6	77.5	80.0
BNPU (Peng et al., 2019)	48.1	77.1	59.2
BERT-ES (Liang et al., 2020)	80.4	67.9	73.7
Conf-MPU (Zhou et al., 2022)	76.6	83.8	80.1
Ours	87.6	76.3	**81.6**
Fully Supervised			
Nooralahzadeh et al. (2019)	92.1	87.9	89.9
Wang et al. (2021)	-	-	90.9
Ours	92.6	91.2	**91.9**

Table 4: Test scores on BC5CDR. The scores of previous methods in the distantly supervised setting are from Zhou et al. (2022).

	ACE2005		
	P	R	F1
Our full model	89.1	89.8	89.5
Shared linear layers	89.3	89.3	89.3
Joint position-span inference	89.4	89.2	89.3
No position-based objectives	88.7	89.9	89.3

Table 5: Test scores of model variants on ACE2005.

B.

	ACE2005								GENIA					
	PER	GPE	ORG	FAC	LOC	VEH	WEA	ALL	Prot.	DNA	CellT.	CellL.	RNA	ALL
S-F1$_{span}$	93.4	91.2	79.7	81.0	78.7	84.8	82.1	89.5	82.9	77.6	74.5	76.3	87.9	80.5
S-F1$_{start}$	93.9	91.2	80.7	81.0	79.0	84.8	82.1	89.9	86.1	80.9	74.5	80.2	88.7	83.2
S-F1$_{end}$	93.9	91.2	81.9	83.1	79.0	86.8	82.1	90.3	87.6	82.6	83.7	82.8	91.0	85.8
L-F1$_{span}$	94.4	91.4	83.0	83.1	79.4	87.2	82.1	90.8	91.6	87.4	84.8	87.2	91.7	89.9

Table 6: Test F1 score breakdowns on ACE2005 and GENIA. Columns compare F1 scores on different entity types. Rows compare F1 scores based on the entire entity span, or only the start or end of entity span. S-F1 denotes the strict F1 requiring the exact boundary match. L-F1 denotes the loose F1 allowing partial overlaps. The color signifies substantially better F1 scores than the corresponding entity span strict F1 scores.

C.

	ACE2005		
	P	R	F1
Dynamic thresholds	89.1	89.8	89.5
Learned global thresholds	88.2	89.0	88.6
Global thresholds tuned on dev	86.3	88.7	87.5

Table 7: Test scores of our method using different thresholding strategies on ACE2005.

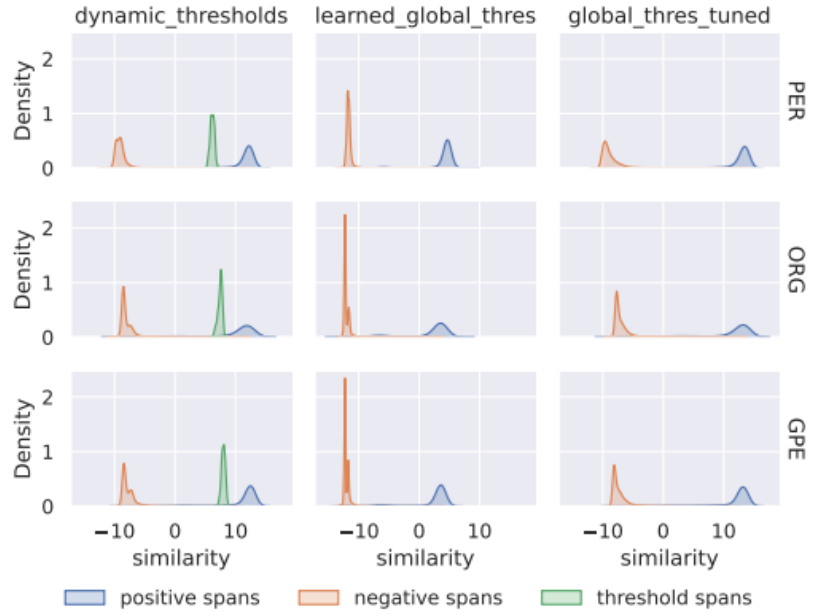


Figure 2: The kernel density estimation of similarity scores between different text spans (entity, non-entity, and threshold spans) and entity types (PER, ORG, GPE) on ACE2005 based on different thresholding strategies.

D.

Error Type	Ent. Type	Predicted ↔ Gold
Modifier Error	VEH	f-14 tomcats (0) ↔ tomcats (0)
	FAC	federal court (0) ↔ court (0)
	VEH	ship (29) ↔ cruise ship (1)
	CellL.	unstimulated T cells (0) ↔ T cells (553)
	Prot.	human GR (0) ↔ GR (88)
	CellT.	lymphocytes (117) ↔ human lymphocytes (18)
	DNA	E6 motif (0) ↔ synthetic E6 motif (0)
Missing Genitive	PER	attendant (3) ↔ attendant's (0)
Annotation Error	PER	Dr. Germ (0) ↔ Dr (1) / Germ (0)
	Prot.	Ag amino acid sequence (0) ↔ Ag (1) / amino acid sequence (6)
	CellL.	EBV-transformed human B cell line SKW6.4 (0) ↔ EBV-transformed human B cell line (0) / SKW6.4 (1)
	DNA	second-site LTR revertants (0) ↔ second-site LTR (0)

Table 8: Examples of common errors among the partially corrected predictions. Red indicates error spans. Blue indicates missing spans. The number after each span mean the span frequency in the training data.