KRISSBERT
Knowledge-Rich Self-Supervision for Biomedical Entity Linking

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao[†], Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon

Microsoft Research

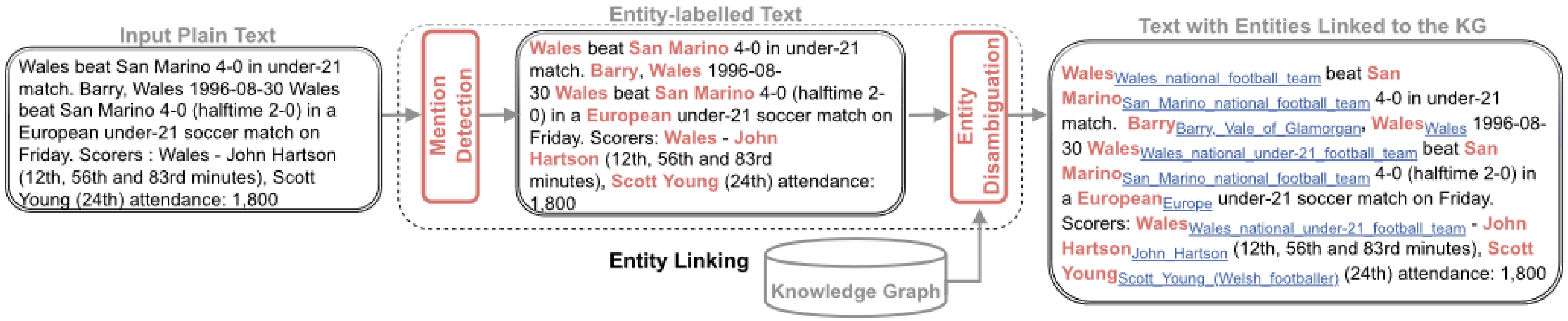
[†]University of Illinois at Urbana-Champaign

Alberto Becerra Tomé

Barcelona, November 2024

Entity Linking in Biomedical Domain

- Process of matching a mention in text to an ontology record.
- Mentions disambiguation is critical in biomedical applications.
- Unified Medical Language System (UMLS): Representative ontology for biomedicine with +4M entities.



Term	Source Vocabulary
Atrial fibrillation	ICD-9-CM
AF	NCI Thesaurus
AFib	MedDRA
Atrial fibrillation (disorder)	SNOMED Clinical Terms
atrium; fibrillation	ICPC2-ICD10 Thesaurus

Image (left): Sevgili, Özge, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. “Neural Entity Linking: A Survey of Models Based on Deep Learning.” *Semantic Web* 13, no. 3 (January 1, 2022): 527–70. https://doi.org/10.3233/SW-222986.
Image (right): https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_001.html

Motivations

- **H1**: Address entity ambiguity and variability in biomedical domain.
- **H2**: Overcome limitations of zero-shot methods such as not considering entity context.
- **H3**: Leverage domain-specific ontologies and self-supervision to enable scalable, accurate biomedical entity linking without labeled examples.

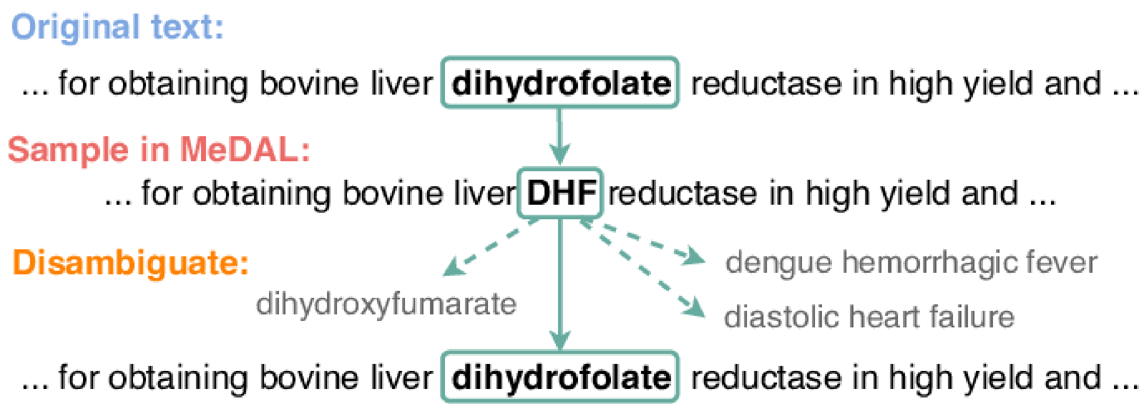
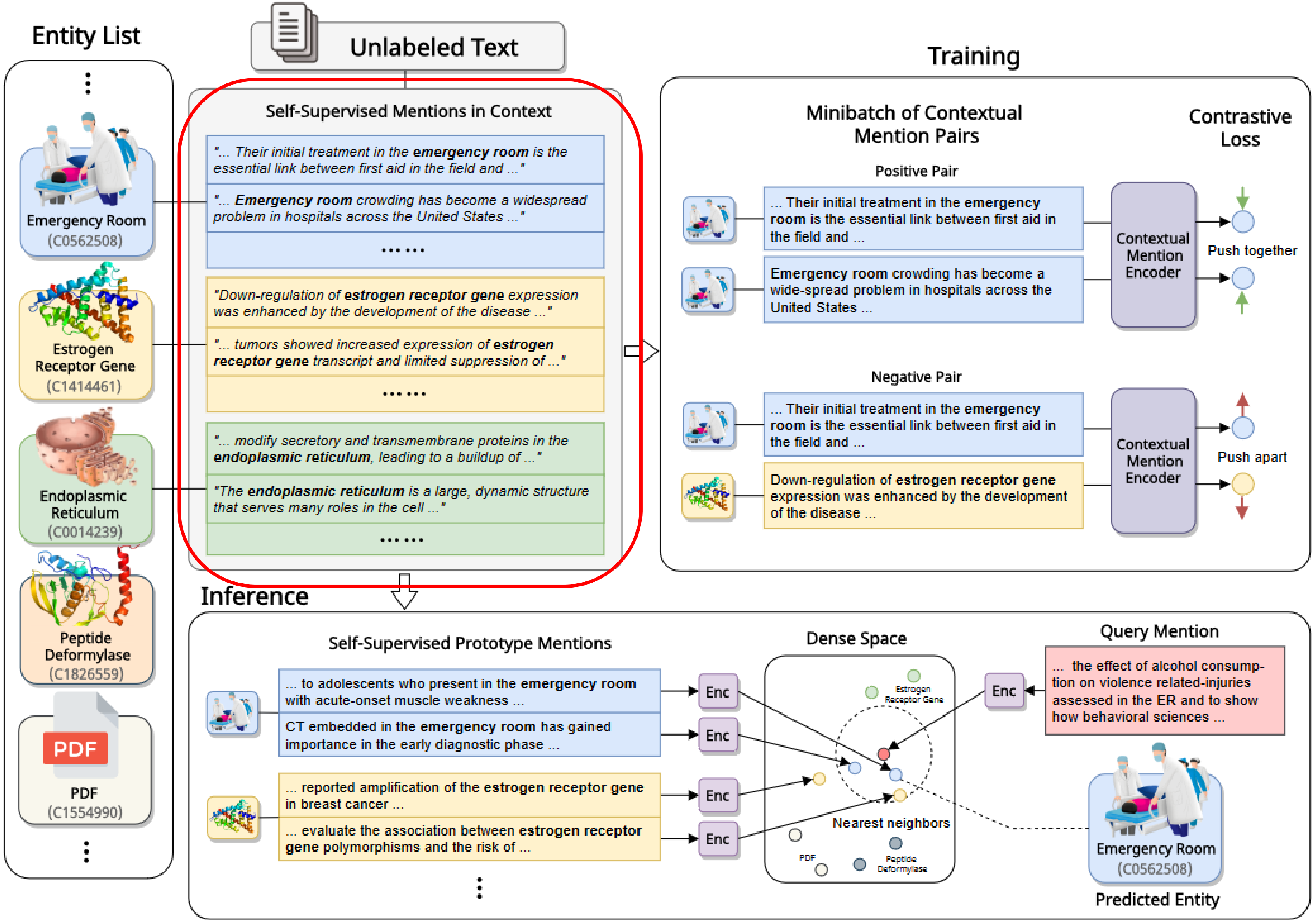


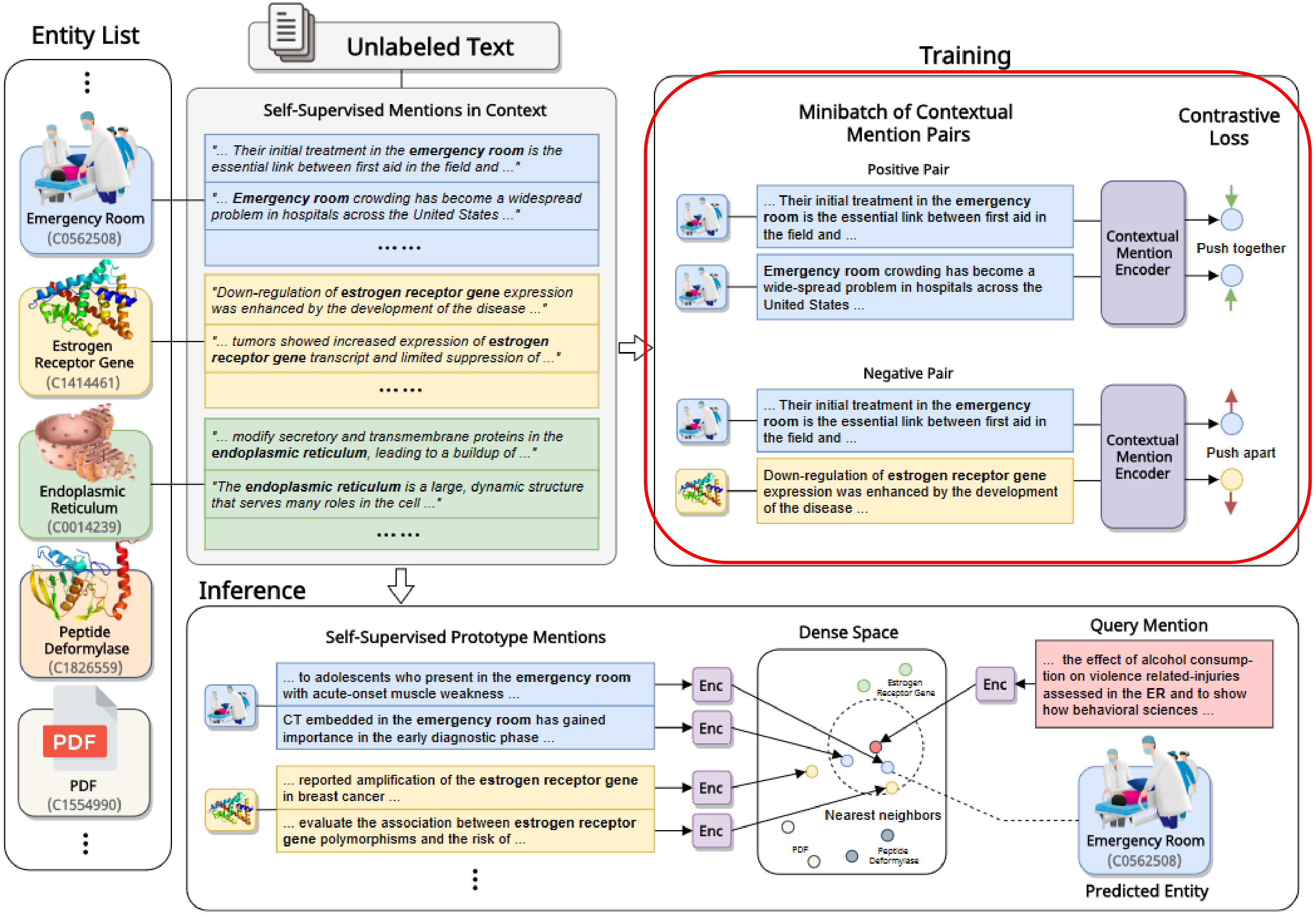
Image: Wen, Zhi & Lu, Xing Han & Reddy, Siva. (2020). MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining. 130-135. 10.18653/v1/2020.clinicalnlp-1.15.

Methodology: Data



- UMLS
  - Over 4M concepts
  - Semantic hierarchy (ISA)
  - Descriptions (6% concepts)
- PubMed
  - Over 1.6B mention examples
  - Dictionary search + context retrieval (64 tokens window)

Methodology: Training



For each minibatch, 2N mentions for N entities are sampled

[CLS] ctx$_L$ [Ms] mention [Me] ctx$_R$ [SEP]

$$\ell_{\mathbf{c}_i,\mathbf{c}_j} = -\log \frac{\exp(\mathbf{c}_i^\top \mathbf{c}_j/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathbf{c}_i^\top \mathbf{c}_k/\tau)}$$

Enriched with N entity-centered references (aliases + semantic):

[CLS] stn [SEP] type [SEP] aliases [SEP]

$$\ell_{\mathbf{c}_i,\mathbf{r}_{e_j}} = -\log \frac{\exp(\mathbf{c}_i^\top \mathbf{r}_{e_j}/\pi)}{\sum_{k=1}^{N} \exp(\mathbf{c}_i^\top \mathbf{r}_{e_k}/\pi)}$$

Methodology: Training

Candidates Generation

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[\ell_{\mathbf{c}_{2k-1},\mathbf{c}_{2k}} + \ell_{\mathbf{c}_{2k},\mathbf{c}_{2k-1}}]$$

$$\mathcal{L}' = \frac{1}{2N}\sum_{k=1}^{N}[\ell_{\mathbf{c}_{2k-1},\mathbf{r}_{e_k}} + \ell_{\mathbf{c}_{2k},\mathbf{r}_{e_k}}]$$

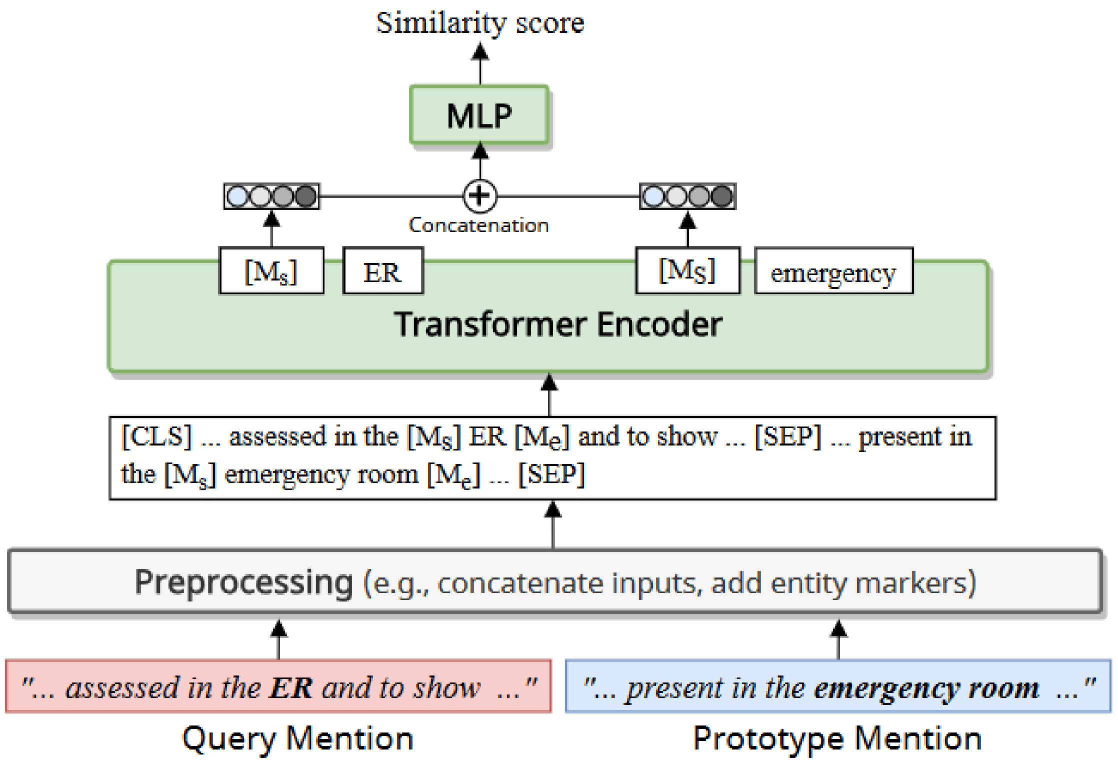
K candidates per mention

Cross-encoder reranker

[CLS] Mention Candidate

Linear Layer

Ranked Candidates



Cross-Encoder Re-ranker

Results: Self-supervision

disease 17,809 | discharge 13,609 | Variety 350k

		NCBI	BC5CDR-d	BC5CDR-c	ShARe	N2C2	MM (full)	MM (st21pv)	Mean
Unsupervised	QuickUMLS	39.7	47.5	34.9	42.1	29.8	12.1	20.0	32.3
	BLINK	49.0	48.7	52.0	32.8	25.1	13.9	19.4	34.4
	SapBERT†	63.0	83.6	96.2	80.4	59.7	37.6	44.2	66.4
	KRISSBERT (self-supervised)	**83.2**±0.5	**85.5**±0.2	**96.5**±0.1	**84.0**±0.1	**67.8**±0.1	**61.4**±0.1	**63.5**±0.1	**77.4**
Supervised	MedLinker	50.5	62.0	80.5	56.8	37.6	32.9	57.6	54.0
	ScispaCy	66.8	64.0	85.3	66.6	54.6	53.1	52.9	63.3
	KRISSBERT (supervised only)	76.9±0.9	85.5±0.7	93.8±0.3	53.9±0.4	29.2±1.2	60.7±0.3	63.7±0.4	66.2
	KRISSBERT (lazy supervised)	**89.9**±0.1	**90.7**±0.1	**96.9**±0.1	**90.4**±0.1	**80.2**±0.1	**70.7**±0.1	**70.6**±0.1	**84.2**

disease 6,892 | disease 5,818 | chemical 4,409 | 21 most common types

Batch size: 512
Learning rate: 10−5
Dropout rate: 0.1
$\mathbf{p}_{\mathbf{mask}} = \mathbf{p}_{\mathbf{replace}} = 0.2$
$\boldsymbol{\tau} = \boldsymbol{\pi} = 1.0$
$\boldsymbol{\alpha} = \boldsymbol{\beta} = 0.5$
K = 100 (number of prototypes for re-ranking)
Initialization weights: PubMedBERT
Training time: 3 hours on 4 NVIDIA V100 GPUs

- Biggest improvement for generic datasets
- Outperforms supervised and unsupervised systems

Results: Lazy Supervision

	KRISSBERT (lazy supervised)	Supervised State of the Art
NCBI	89.9	89.1 (Ji et al., 2020)
BC5CDR	93.7	91.3 (Angell et al., 2021)
ShARe	90.4	91.1 (Ji et al., 2020)
N2C2	80.2	81.6 (Xu et al., 2020)
MM (full)	70.7	45.3† (Mohan and Li, 2019)
MM (st21pv)	70.6	74.1 (Angell et al., 2021)

Table 6: Comparison of test accuracy of KRISSBERT with lazy learning (§3.7) and supervised state of the art. †Prior work generally avoids evaluating on the full MM dataset; we can only find one published result which combines boundary detection and linking.

Results: Disambiguation

	Ambiguous(%)	SapBERT	KRISSBERT
NCBI	43.2	57.1	64.5
BC5CDR-d	30.7	63.9	64.5
BC5CDR-c	11.5	76.4	76.5
ShARe	48.5	67.5	72.4
N2C2	67.5	50.7	58.2
MM (full)	67.8	24.8	48.9
MM (st21pv)	69.4	29.6	52.5

Table 5: Accuracy comparison on ambiguous cases.

Example: "... *5 days of oral prednisone in treatment of adults with **mild** to moderate asthma exacerbations ...*"
SapBERT prediction: surface form MILD, which is shared by multiple entities, such as Mild Severity of Illness Code (C1547225), Mild Adverse Event (C1513302).
KRISSBERT prediction: Mild asthma (C0581124)
KRISSBERT predicted prototype:
"***Mild asthma*** *exacerbations in a group of children with cough as a dominant symptom ...*"

*We consider a mention as ambiguous if it can't be matched to a unique entity as is.

Results: Design

	NCBI	BC5CDR-d	BC5CDR-c	ShARe	N2C2	MM (full)	MM (st21pv)	Mean
KRISSBERT	**83.2**	**85.5**	**96.5**	**84.0**	**67.8**	**61.4**	**63.5**	**77.4**
— cross-attention re-ranking	82.8	85.0	95.1	83.4	65.0	59.4	61.3	76.0
— mention pair contrast	77.9	82.2	93.3	75.0	56.3	47.8	49.9	68.9
— aliases	83.2	85.2	96.4	84.0	67.7	61.0	63.2	77.2
— semantic hierarchy	82.7	85.1	96.4	83.0	65.7	59.0	61.5	76.3
— entity description	83.1	85.4	96.3	84.0	67.8	61.2	63.4	77.3
Initialize w. BERT	79.3	80.6	94.4	74.5	58.4	53.9	55.3	70.9

Conclusions

H1: Address entity ambiguity and variability in biomedical domain.

C1: Consistent performance even in MedMentions (high variability, large scale)

H2: Overcome limitations of zero-shot methods such as not considering entity context.

C2: Scalable and efficient context encoding and entity-centered encoded mentions (Design).

H3: Leverage domain-specific ontologies and self-supervision to enable scalable, accurate biomedical entity linking without labeled examples.

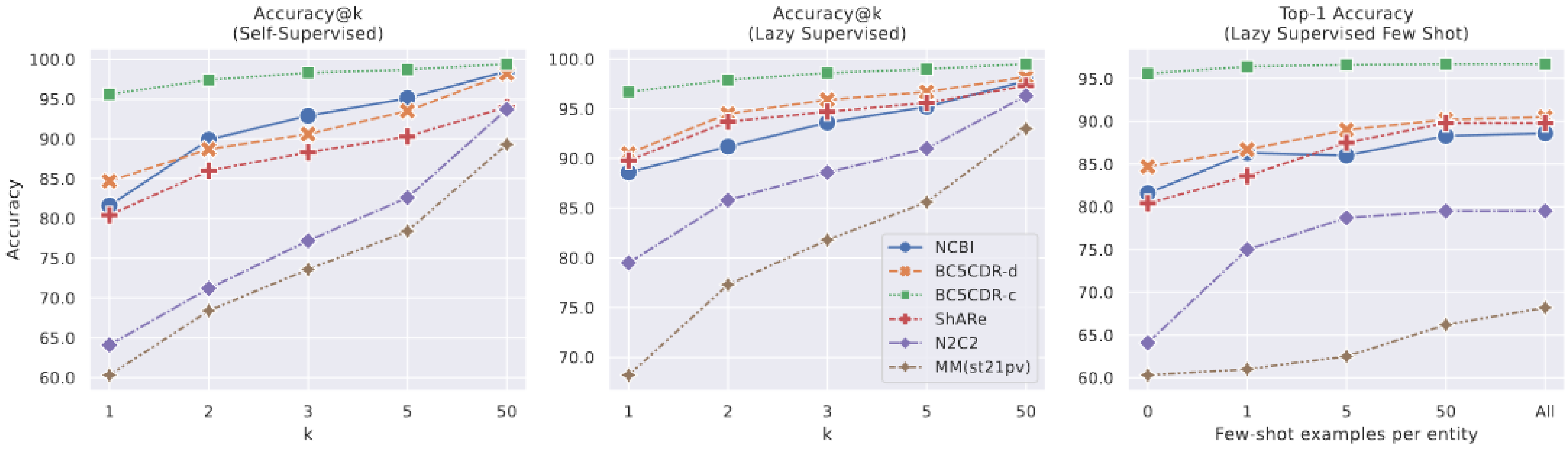
C3: Self-supervised inference consistently outperforms most of the SOTA systems including SapBERT.

Promising Tool for Low-Resource Settings: Valuable solution for biomedical NLP tasks where labeled data is scarce.

Lazy-supervision Technique: Verification of positive impact of enriching self-supervised mentions with gold standard.

Limitations and Proposed Solutions

- Huge difference in Accuracy@k in some datasets.
  - **Solution**:
    - Initialize cross-encoder re-ranker with strong pretrained weights using models such as jina-reranker-v2.
    - Using more advanced techniques for negative mining in contrastive learning such as Top-k with percentage to positive threshold [3].

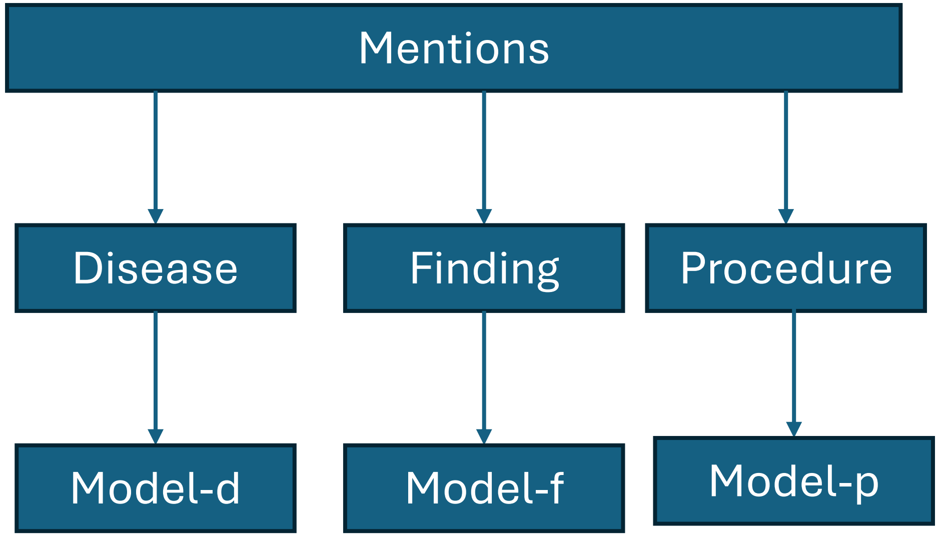


Limitations and Proposed Solutions

- Confusing entities in different ontology branches such as Procedures and Substances
  - **Solution:** Split the ontology in different types + NERC + Specific fine-tuned model

Mention: "... *NTeff cells appeared to have lower* ***expression*** *of Foxp1* ..."
Gold entity: Protein Expression (C1171362)
KRISSBERT **prediction**: Expression Procedure (C0185117)
KRISSBERT **predicted prototype**: "... ***expression*** *of a myeloid differentiation antigen, Mo1* ..."

Mention: "... *On admission included BUN /* ***creatinine*** *of 33/2.1 . Sodium 141 .* ..."
Gold entity: Creatinine Measurement (C0201975)
KRISSBERT **prediction**: Creatinine (C0010294)
KRISSBERT **predicted prototype**: "... *Sorbent binding of urea and* ***creatinine*** *in a Roux-Y intestinal segment.* ..."



References

[1] Zhang, Sheng, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. “Knowledge-Rich Self-Supervision for Biomedical Entity Linking.” arXiv, May 23, 2022. https://doi.org/10.48550/arXiv.2112.07887.

[2] Sevgili, Özge, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. “Neural Entity Linking: A Survey of Models Based on Deep Learning.” Semantic Web 13, no. 3 (January 1, 2022): 527–70. https://doi.org/10.3233/SW-222986.

[3] Moreira, Gabriel de Souza P., Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. “NV-Retriever: Improving Text Embedding Models with Effective Hard-Negative Mining.” arXiv, July 22, 2024. https://doi.org/10.48550/arXiv.2407.15831.

[4] Zhang, Sheng, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. “Knowledge-Rich Self-Supervision for Biomedical Entity Linking.” arXiv, May 23, 2022. https://doi.org/10.48550/arXiv.2112.07887.

[5] Gallego, Fernando, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J. Veredas. “ClinLinker: Medical Entity Linking of Clinical Concept Mentions in Spanish.” arXiv, April 9, 2024. https://doi.org/10.48550/arXiv.2404.06367.

[6] Liu, Fangyu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. “Self-Alignment Pretraining for Biomedical Entity Representations.” arXiv, April 7, 2021. https://doi.org/10.48550/arXiv.2010.11784.

Thank You